



# QUANTUM FIELD THEORY

## A 20th Century Profile

With a Foreword by Freeman J. Dyson

Edited By Asoke N. Mitra



**HINDUSTAN  
BOOK AGENCY**



**INDIAN NATIONAL  
SCIENCE ACADEMY**





---

# Quantum Field Theory

*A Twentieth Century Profile*

---

With a Foreword by Freeman J Dyson

Edited By

ASOKE N MITRA

Formerly INSA-Einstein Professor  
Indian National Science Academy

 **HINDUSTAN  
BOOK AGENCY**



**INDIAN NATIONAL  
SCIENCE ACADEMY**

Published Jointly by Hindustan Book Agency (India) and  
Indian National Science Academy.

Copyright © 2000 by Hindustan Book Agency (India) and  
Indian National Science Academy.

No part of the material protected by this copyright notice  
may be reproduced or utilized in any form or by any means,  
electronic or mechanical, including photocopying, record-  
ing or by any information storage and retrieval system,  
without written permission from the copyright owner, who  
also has the sole right to grant licences for translation into  
other languages and publication thereof.

All export rights for this edition vest exclusively with  
Hindustan Book Agency (India). Unauthorized export is a  
violation of Copyright Law and is subject to legal action.

*INSA Editor of Publications:*

Professor S. K. Malik

*INSA Editorial Staff:*

J. Saketharaman, AES-I

Rajan Phull, SO-I

ISBN 81-85931-25-9

Printed by Chaman Enterprises, Delhi

## DEDICATION

The BOOK is dedicated to **three** great Teachers:

**Jatindranath Mitra** who introduced his son to Mathematical Physics;

**Hans Bethe** who showed the dual virtues of *simplicity* and *thoroughness* in Physics;

**Freeman Dyson** who gave a glimpse of the *beauty* that Physics exudes  
when clothed in the language of Mathematics.



### INSA PRESIDENT'S NOTE

To mark the International Mathematics Year 2000, the Indian National Science Academy had decided to publish three specialized publications on the following subjects: Number Theory, Non-linear Phenomenon and Quantum Field Theory. The Academy had invited one of its Fellows, Dr. Asoke N. Mitra, former Einstein Professor of Indian National Science Academy, to prepare a monograph on a suitable theme and solicit articles from internationally acknowledged experts. The result of his endeavour is a comprehensive volume on *Quantum Field Theory – A Twentieth Century Profile*. The foreword to this volume is by Freeman Dyson, one of the Founding Fathers of modern Quantum Field Theory.

The Academy thanks Dr. Mitra for his efforts and all the distinguished contributors to the book for their ready response.

I very much hope that the present monograph will be a valuable addition to the literature on Quantum Field Theory and will be appreciated by the Mathematical Physics Community at large.

New Delhi  
8th February 2000

Goverdhan Mehta  
President, INSA





# Foreword

Freeman Dyson \*

Institute for Advanced Study, Princeton, New Jersey, USA

## The Impact of Field Theory on Physics of the Twentieth Century

I am proud to count Dr. Asoke Mitra, the editor of this volume, among my students. After serving his apprenticeship as a field theorist at Cornell University in the USA, he chose to return home to help build up science in India. The choice was not easy, at a time when science in America was flourishing and rapidly pushing ahead, while science in India was struggling to overcome the obstacles imposed by geography and history. Dr. Mitra sacrificed his chance of a brilliant research career in America, in order to serve his country and his people. I deeply respect that choice, and I rejoice that his sacrifice was not made in vain. After a fruitful career as a pioneer and teacher of modern science in India, he now stands at the center of the vibrant scientific community that he helped to create. This volume is, among other things, a monument to his vision.

Quantum field theory was a good subject, and remains a good subject, for cultivation in India. Homi Bhabha was one of its early practitioners. It is firmly rooted in experimental science, but it can be cultivated without elaborate apparatus. It is well suited to a country with limited material resources and an ample supply of talented people. Indian field-theorists are well placed to play a leading role in bringing modern science to Asia. Anywhere in the world, a group of quantum field theorists can make serious contributions to science, if they have a modest computer with connections to the internet and the world wide web. Especially in the last ten years, since the rapid development of string theory brought new excitement to the field, quantum field theory has become a world-wide enterprise. Everywhere you go, wherever there are scientists and students, you find string-theorists.

When I was growing up sixty years ago, I used to read books by the writer Peter Fleming describing his travels in remote places. He wrote a book, "News from Tartary", about his travels in Central Asia. At that time the city of Urumchi, in the mountains of north-west China, was the most inaccessible place on earth. All the trails to Urumchi were blocked by rival war-lords and bandits. Fleming tried for several years to reach Urumchi but never succeeded. I thought of Urumchi then as a legendary place, for ever beyond the reach of civilization. I was not sure that it really existed. Now, sixty years later, I received a post-card from Urumchi, sent by Andrew Strominger, an American string-theorist who happened to be visiting

---

\*Email: [dyson@ans.ias.edu](mailto:dyson@ans.ias.edu)

there. The post-card has a picture of magnificent snow-covered mountains on the front. On the back Strominger wrote, "There is a lively group of string-theorists here". Urumchi is now a thriving industrial city, called Wulumuchi by the Chinese. Quantum field theory penetrates all barriers.

Quantum field theory was invented by Europeans in the nineteen-twenties, very soon after they had invented quantum mechanics. The first quantum field theory was the quantum electrodynamics of Dirac, describing the emission and absorption of electromagnetic radiation by atoms. At that stage, only the Maxwell field was quantized. Dirac's theory was immediately successful in explaining the observed behavior of radiation. It agreed with all known experiments in atomic physics, and with Planck's law of black-body radiation. Already in 1927 it was clear that quantum electrodynamics was basically correct. Within the next two years, Dirac's theory was extended and improved by Fermi, Heisenberg, Pauli, Jordan and Wigner, so that it included all the known particles and interactions. The framework of quantum field theory was then complete. It appeared to give a correct description of all physical phenomena with the exception of gravitation.

Confidence in quantum field theory rested on three foundations. First, the part of it that concerned electromagnetic processes was confirmed by accurate experiments. Second, the theory as a whole was mathematically elegant, unifying in a beautiful way the principles of special relativity and quantum mechanics. Third, the theory explained the most striking fact about elementary particles, the fact that all particles belong to a small number of species, with the particles within each species indistinguishable. All electrons are indistinguishable, with the same mass and charge, because they are all embodiments of the same field. A few years later the positron was discovered, with the same mass as the electron but with equal and opposite charge. This was final confirmation that matter is built out of quantum fields, each charged field giving rise to particles and anti-particles linked by a deep underlying symmetry. Quantum field theory explained the existence of anti-particles and also explained the symmetry.

The basic framework of quantum field theory has remained the same from the nineteen-twenties until the present. But the details of the theory have changed as new kinds of fields and particles have been discovered. The most important change was the introduction of new gauge fields, in addition to the gauge field of electrodynamics, to describe the weak and strong interactions. The Weinberg-Salam model of the weak interactions introduced three gauge fields, each associated with a massive new species of particle. When the W and Z particles predicted by the model were discovered, nobody could any longer doubt that quantum field theory was correct. A few years later, quantum chromodynamics provided a model of the strong interactions within the same framework. The new gauge fields, incorporating the weak and strong interactions, give rise to the rich diversity of phenomena that are seen in high-energy experiments.

All through its history, quantum field theory has had two faces, one looking outward, the other looking inward. The outward face looks at nature and gives us numbers that we can calculate and compare with experiment. The inward face looks

at mathematical concepts and searches for a consistent foundation on which to build the theory. The outward face shows us a brilliantly successful theory, bringing order to the chaos of particle interactions, predicting experimental results with astonishing precision. The inward face shows us a deep mystery. After seventy years of searching, we have found no consistent mathematical basis for the theory. When we try to impose the rigorous standards of pure mathematics, the theory becomes undefined or inconsistent. From the point of view of a pure mathematician, the theory does not exist. This is the great unresolved paradox of quantum field theory.

To resolve the paradox, during the last twenty years, quantum field theorists have become string-theorists. String theory is a new version of quantum field theory, exploring the mathematical foundations more deeply and entering a new world of multidimensional geometry. String theory also brings gravitation into the picture, and thereby unifies quantum field theory with general relativity. String theory has already led to important advances in pure mathematics. It has not yet led to any physical predictions that can be tested by experiment. We do not know whether string theory is a true description of nature. All we know is that it is a rich treasure of new mathematics, with an enticing promise of new physics. During the coming century, string theory will be intensively developed, and, if we are lucky, tested by experiment. We can rejoice that Indian scientists will be active participants in this adventure.



## Preface

On the eve of the International Mathematics Year 2000, the Indian National Science Academy had decided to bring out some selected volumes on different facets of Physics and Mathematics. The basic idea was somewhat on the lines of the American Physical Society's plans to bring out its souvenir Volume "More Things On Heaven And Earth", at the end of the Millenium (which also coincided with its Centenary Year). However the Academy's plans were much more modest, being limited to a few topical aspects of these subjects within the scope of its journal specialization. The present volume in particular aims to highlight the impact of Field Theory on the evolution of Physics of the Twentieth Century. The choice of Field Theory as a central theme for this particular Volume was dictated by the consideration that, as the single most important concept in Physics to be discovered in this century, it would register a commanding presence on a vast array of topics which would bear standing testimony to the success story of this unique armour in the arsenal of Physics. We were indeed fortunate to receive an endorsement of the QFT theme from Freeman Dyson, one of the architects of modern Field Theory, in a Foreword to this Volume, under the title "The Impact of Field Theory on Physics of the Twentieth Century" in which he has extolled the universal appeal of Field Theory in Physics. This theme is projected in this Book through a selection of articles (by acknowledged experts) in those areas where the impact of QFT has been especially pronounced, from particle physics to string theory (with several interpolating stages of development), and extending to some facets of astrophysics and the physics of condensed matter.

The emphasis in the Book is mainly on quantum field theory (QFT), so the standard Einstein's (geometric) theory of gravitation (whose quantum formulation is still a distant goal), is not in its direct purview. Of course the String Theory route for evolution of QFT (which holds the key to a potential understanding of gravitation) has been included, but not in a specialized enough form to do proper justice to gravitation. The actual contents of the Book are of necessity governed by the access to acknowledged experts in their respective areas of expertise within a relatively short span of time, yet the response has been quite encouraging.

In the area of particle physics the emphasis is mainly on symmetries, topologies, gauge theories and renormalization groups. While electroweak interactions have been treated with standard rigour, the strong interaction sector has needed greater filtration, so as to conform to the basically QFT-oriented thrust of the Book.

A distinct feature of this Book (not usual for an edited book) is that its theme has been highlighted through a comprehensive Editorial Summary of all the 33 articles, preceding their classified presentation in *six* distinct parts:

A) basic structure of QFT; B) topological aspects of QFT; C) miscellaneous formal methods in QFT;  
D) extension of QFT frontiers; E) QFT in  $2 + 1$  dimensions; and F) strong interaction methods in QFT/QCD.

The contributors range from veterans like (the late) Vladimir Gribov, Marcos Moshinsky, Kazuhiko Nishijima, John Schwarz, Dmitri Shirkov and Edward Witten, to a string of acknowledged experts in their respective fields of expertise, all the way to a few young and promising workers. The Book concludes with a modern perspective on the observable limitations of Quantum Field Theory.



The Book contains two Articles not directly written by the authors concerned in response to the Academy's invitation, but nevertheless central to the QFT theme. The first is a set of three (unpublished) Orsay Lectures on Confinement by the late Vladimir Gribov during 1992-93 (before his death), as compiled by his associates Profs. Dokhshitzer, Ewarz and especially Julia Nyiri (Mrs. Gribov). We are grateful to all of them for permission to reproduce these notes in this Volume, especially because the Academy has nostalgic memories of Prof Gribov's close association with the Academy during the Sixties. We are particularly indebted to Prof. Olivier Pene for his invaluable help in making these Lectures available to us in the form of LPTHE preprints and LANL hep-ph documents. The second Article is an outstanding paper entitled "Quantum Field Theory and the Jones Polynomial" by Edward Witten, originally published in Commun.Math.Phys. **121**, 351-399 (1989). We are extremely grateful to Prof. Witten as well as the Publishers (Springer-Verlag, Germany) for permission to reproduce this paper in the present Volume in exactly the same format in which it had appeared in the original Journal. Both the Gribov and the Witten Articles appear in Part A of the Book.

The wide range of topics covered makes the Book more than just an introductory text book on QFT. It is recommended as a reference book for a broad spectrum of readership, from fresh post docs in most key areas of QFT to the specialists in evolving areas. And the freelance researcher in QFT should find enough "appetizers" among the contents to kindle his interest in the field. However while no efforts have been spared to maintain a basically field-theoretic texture for the diverse topics covered in the Book, no uniformity can be claimed for their mathematical standards of presentation, reflecting as they do the variations in the state of the art in the respective fields.

The Academy is grateful to all the contributors for their timely cooperation with their respective Articles. Above all, the Editor is deeply indebted to his Teacher and Mentor, Freeman Dyson, for readily consenting to give a Foreword to this Volume which is the first such venture of the Indian National Science Academy.

The Book has been composed directly under the auspices of the Academy under the orders of the President, Prof. Goverdhan Mehta. We are grateful to Prof. Mehta for his personal interest in this venture, as well as to Prof S.K.Malik (Editor of the Academy) for his unstinted help in getting the process going. And thanks to the active interest of the Executive Secretary (Mr. S.K.Sahni), the editorial staff, especially Mr. Saketaraman and Mr. Rajan Phull in the final stages of production. We are especially grateful to Mr. Jayan without whose deep involvement, the crucial LaTeX processing of the Book would not have been possible. Thanks are also due to Mr. Santosh Malik, for editorial help. The technical advice of Dr.Vineet Ghildyal (of the Delhi University Library System) in the specialized aspects of composition in LaTeX format, has been truly invaluable, and is most gratefully acknowledged.

Asoke N Mitra

# CONTENTS

Foreword by Freeman Dyson	vii
Preface	xi
<i>Editor's Summary</i> : Dimensions Of Field Theory – From Particles To Strings by A.N.Mitra	3–21
 <b>Part A : Basic Structure Of QFT</b>	 <b>22</b>
1. <i>D.V.Shirkov</i> : Evolution Of The Bogoliubov Renormalization Group	25–58
2. <i>S.Szipigel and R.J.Perry</i> : The Similarity Renormalization Group	59–81
3. <i>V.Novikov</i> : Quantum Field Theory And The Standard Model - Bird's Eye View	82–113
4. <i>P.K.Kabir</i> : Broken Reflection Symmetries	114–121
5. <i>D.Boyanovsky and H.J.de Vega</i> : Dynamics Of Symmetry Breaking Out Of Equilibrium - From Condensed Matter To QCD And The Early Universe	122–157
6. <i>Vladimir.N.Gribov (Late)</i> : Orsay Lectures On Confinement	
(I) hep-ph/9403218	158–180
(II) hep-ph/9404332	181–187
(III) hep-ph/9905285	188–199
7. <i>K. Nishijima and M.Chaichian</i> : An Essay On Color Confinement	200–207
 <b>Part B : Topological Aspects Of QFT</b>	 <b>209</b>
8. <i>R.Kaul</i> : Topological Quantum Field Theories - A Meeting Ground For Physicists and Mathematicians	211–232
9. <i>Edward Witten</i> : Quantum Field Theory And The Jones Polynomial	233–281
10. <i>H.Banerjee</i> : Chiral Anomalies In Field Theories	282–296
11. <i>Wei-Min Zhang</i> : Coherent States In Field Theory	297–323
12. <i>N.Mukunda</i> : Pancharatnam, Bargmann And Berry Phases - A Retrospective	324–336
13. <i>J.Schechter and H.Weigel</i> : The Skyrme Model For Baryons	337–369
 <b>Part C : Formal Methods In QFT</b>	 <b>371</b>
14. <i>R.Ramanathan</i> : Euclidean Methods In Quantum Field Theory	373–382
15. <i>Ashok Das</i> : Topics In Finite Temperature Field Theory	383–411
16. <i>B.M.Sodermark</i> : Integrable Models And The Toda Lattice Hierarchy	412–436

17. <i>Prem P Srivastava</i> : Perspectives Of Light-Front Quantized Field Theory -Some New Results	437 478
18. <i>D.S.Kulshreshtha</i> : Gauge Symmetry In Chiral Electrodynamics	479 489
19. <i>L.Lusanna</i> : Towards A Unified Description Of The Four Interactions In Terms Of Dirac-Bergmann Observables	490 518
<b>Part D : Extension Of QFT Frontiers</b>	<b>519</b>
20. <i>R.N.Mohapatra</i> : Supersymmetry And Particle Physics	521 543
21. <i>N.Sakai</i> : Supersymmetry In Field Theory	544 570
22. <i>Werner Nahm</i> : Conformal Field Theory: A Bridge Over Troubled Waters	571 604
23. <i>John H Schwarz</i> : Superstring Theory - An Overview	605 610
24. <i>J.Maharana</i> : Recent Developments In String Theory	611 664
25. <i>L.Bonora</i> : Yang-Mills Theory And Matrix String Theory	665 699
<b>Part E : QFT In 2 + 1 Dimensions</b>	<b>701</b>
26. <i>Avinash Khare</i> : Fractional Statistics And Chern-Simons Field Theory In 2 + 1 Dimensions	702 729
27. <i>R.Rajaraman</i> : Chern Simons Field And Composite Bosons In The Quantum Hall System	730 741
<b>Part F : Methods Of Strong Interactions In QFT</b>	<b>743</b>
28. <i>Olivier Pene</i> : Hadrons From QCD - Achievements And Prospects	745 753
29. <i>L.S.Kisslinger</i> : QCD Sum Rules In Hadronic And Nuclear Physics	754 794
30. <i>V.A.Karmanov</i> : Light-Front Dynamics	795 825
31. <i>A.N.Mitra</i> : 3D-4D Interlinkage Of B-S Amplitudes - Unified View Of $Q\bar{Q}$ and $QQQ$ Dynamics	826 856
32. <i>Marcos Moshinsky</i> : The Harmonic Oscillator In Quantum Theory - A Powerful Bridge In Physics	857 874
<b>Conclusion</b>	<b>875</b>
33. <i>D.Home</i> : Modern Perspectives On Foundations Of Quantum Mechanics	877 900

# The Dimensions Of Field Theory – From Particles To Strings \*

A.N.Mitra †

## Abstract

This is an editorial summary of the contents of a Book comprising a set of Articles by acknowledged experts dealing with the impact of Field Theory on major areas of physics (from elementary particles through condensed matter to strings), arranged subjectwise under six broad heads. The Book which emphasizes the conceptual, logical and formal aspects of the state of the art in these respective fields, carries a Foreword by Freeman Dyson, and is to be published by the Indian National Science Academy on the occasion of the International Mathematical Year 2000. The authors and full titles of all the Articles (33) are listed sequentially (in the order of their first appearance in the narration) under the bibliography at the end of this Summary, while a few of the individual articles to appear in the Book are already available on the LANL internet.

## 1 Birth, Decline And Rebirth Of Field Theory

If one must choose one single item of Twentieth Century Physics which stands out by the yardstick of most pervasive and decisive influence on its total development, Quantum Field Theory (QFT) certainly wins hands down. Historically, QFT was born out of the marriage of Relativity and Quantum Theory, at a hefty price of mathematical self-consistency underlying the celebrated Dirac Theory, whose full significance took several stages to unfold through the vicissitudes of logical deduction, greater significance from the conceptual point of view, was the realization that the "sea of negative energy states" was already a tacit admission of the failure of relativistic quantum mechanics of a single particle, in favour of a collective many-particle, or *field* description, a fact which was to be driven home by Dyson in his Cornell lectures of 1952. And once this realization dawned on the pioneers, the Klein-Gordon theory of scalar particles found a natural place in the new scenario, at the hands of Pauli-Weisskopf(1934) who now found little difficulty in quantizing these bosonic particles just as easily as the Dirac theory had done to fermions. Thus was born "Quantum Field Theory" (QFT) in its full glory, with Anti-matter playing a symmetrical role to Matter, irrespective of its fermionic or bosonic nature. [ Feynman's brilliant positron theory was a bold attempt to resurrect the single particle quantum mechanics description via "zigzag" diagrams (negative time propagation of negative energy electrons), but the more universal language of Field Theory eventually carried the day].

QFT registered its first major success in the Covariant formulation of QED at the hands of Tomonaga and Schwinger on the one hand, and Feynman on the other, with Dyson playing the catalyst-role in synthesizing the two. This theory, in the course of circumventing unphysical infinities in the measurable quantities, gave rise to a new dogma of *Renormalizability* which was to act as the yardstick of acceptability of theories to come. This dogma, together with the independent principle of "Gauge Invariance" (already in-built in QED), were to be two pillars of QFT in its march towards greater victories to come, especially in the formulation of strong interaction theories on analogous lines to QED. This led to the Yang-Mills theory (1954) of  $SU(2)$

---

\*Editorial Summary : hep-ph/9911450

†Email: (1) ganmitra@nde.vsnl.net.in ; (2) anmitra@csec.ernet.in

put both aspects together and derived the "RG-equations" in a form which brings out the 'scaling' properties of the electron and photon propagators. Thus RG invariance boils down to the invariance of a solution w.r.t. the manner of its parametrization. These equations were further developed and made more rigorous with mathematicians and physicists working in tandem, so that renormalization became a well-developed method at the computational level. But the underlying physical concepts behind these equations took some more time to unfold until after Kadanov's, and especially Wilson's pioneering work on the understanding of the "critical indices" in phase transitions brought out the real physics behind the RG equations.

Wilson's work revealed the rich applicational potential of the RG ideas in various fields of physics, from 'critical phenomena' (spin lattices, polymer theory, turbulence) in condensed matter physics, to QCD parameters like the strong coupling constant  $\alpha_s$  and the 'running mass'  $m(p^2)$ . In particular, the discovery of Asymptotic freedom in QCD allowed physicists to produce a logically consistent picture of renormalization, one in which the perturbative expansions at any high energy scale can be matched with one another, without any need to deal with intermediate expansions in powers of a large coupling constant. Another important aspect of these RG equations which has been emphasized by the Dubna School, is the concept of *functional self-similarity* in mathematical physics, which has led to applications like the study of strong non-linear regimes: asymptotic behaviour of systems described by non-linear partial differential equations; problem of generating higher harmonics in plasmas, and so on. The Book begins with a perspective Article by **Dmitri Shirkov**[1] on all aspects of the subject, from an introduction to RG in QFT to an overview of its methodology, together with applications of RG ideas in some important arenas of physics.

A relatively new approach to RG theory, termed "Similarity Renormalization Group" (SRG) was launched in this decade by Wilson and Glazek, as well as Wegner, and is based on the perception that divergences are in the first place due to the *locality* of the primary interactions. For a proper understanding of the features of the SRG theory, it is enough to consider only the non-relativistic quantum mechanics (the usual UV divergences of relativistic QFT are not relevant here!), where the locality condition on the potentials at all scales corresponds to taking only delta functions and their derivatives. The associated divergences can be regulated by introducing cut-offs whose effects may be removed by renormalization.

In the SRG, the transformations that explicitly "run" the cut-off parameter are developed. These similarity transformations are of course unitary, and constitute the group elements of SRG. They are characterized by a "running" cut-off on energy differences (not states). If the Hamiltonian is viewed as a large matrix, these cut-offs limit the off-diagonal matrix elements, and as they are gradually reduced, the Hamiltonian is forced towards the diagonal form. The perturbation expansion of the transformed Hamiltonians contains no small energy denominators, so that the expansion does not break down unless the strengths of the interactions themselves are large. With the help of an associated concept of *coupling coherence*, SRG acquires respectability as a proper theory with the *same* number of parameters as the original (fundamental) theory. A review of the formalism and working of SRG is given by **R J Perry**[2], using as an example the exactly soluble case of a simple 2D delta function to act as a laboratory for testing the convergence of the SRG method in some detail.

## 2.2 Standard Model And Electroweak Coupling

The Gauge Principle, as a central ingredient of QFT, needed to be supplemented with fresh ideas and paradigms, within its broad framework, to extend its tentacles further. One such idea was based on the degenerate structure of the vacuum, dominated by vales and hills, which crystallized eventually as a new theme termed "Spontaneous Symmetry Breaking" (*SSB*), together with its companion "Dynamical Breaking of Chiral Symmetry" (*DB $\chi$ S*), which would now enable gauge fields to acquire mass in a subtle but self-consistent manner. Armed with this paradigm, the Gauge Theory registered a signal success in the Weak interaction sector, culminating in the Glashow-Salam-Weinberg (*GSW*) Model of Electro-weak Interactions, which offered a unified view of weak and electromagnetic interactions in the form of an  $SU(2) \otimes U(1)$  gauge theory. A more ambitious form of unification of the three principal gauge fields as a straightforward extension of the *GSW*, so

as also to include the strong ( $GUT$ ), did not unfortunately bear fruit, so that, for the time being, the "Standard Model" ( $SM$ ) has had to rest content with only a partial unification  $SU(3) \otimes SU(2) \otimes U(1)$  of these gauge fields. Nevertheless this episode brings out a truism about the unpredictability of Nature, viz., its refusal to yield to a particular strategy for a second time, merely on the strength of its success on a previous occasion.

In a highly instructive and self-contained Article, V Novikov [3] gives a panoramic view of the conceptual and methodological framework of QFT (with the ingredients of gauge principle, renormalization group, and spontaneous symmetry breaking) that have been employed in the formulation of  $SM$  for elementary particle physics. He dwells in particular on the Higgs mechanism for the generation of the fermion masses for several generations, and brings out the powers of "loop corrections" in  $SM$  to predict accurate bounds on the masses of as yet undiscovered particles. This is vividly illustrated by the "correct" mass of the  $t$ (*op*)-quark *ahead* of its experimental discovery, stringent limits on the Higgs mass from the "Landau pole" structure of the running coupling constant, and the windows to the "physics beyond  $SM$ " that such analyses provide.

### 2.2.1 Discrete Symmetries in SM

An essential aspect of the Standard Model concerns the role of discrete symmetries  $P, C, T$  in determining the structure of the electroweak coupling. This subject has had a long history since the original Lee-Yang discovery of  $P$ -violation, going through successive phases of chiral symmetry (Landau-Salam),  $CP$  invariance (Lee-Oehme-Yang), its subsequent violation (Cronin-Fitch), and *ipso facto* (?)  $T$ -violation, a topic of intense experimental activity today. [This last is of course an immediate consequence of  $TCP$ -invariance (Pauli-Lueders Theorem), which puts the existence of antiparticles exactly on par with particles]. A brief state-of-the-art review of the subject by P K Kabir [4] follows.

## 2.3 Dynamics Of Symmetry Breaking

Just as "Symmetry dictates interactions" - (C.N.Yang at the First Asia Pacific Conf, Singapore, 1983), the dynamical effects of its *breaking* (whether spontaneously or dynamically) during out-of-equilibrium phase transitions is equally at the root of a whole range of phenomena from condensed matter to particle physics, and so on, all the way to early universe cosmology. Indeed the dynamics of non-equilibrium phase transitions and the *ordering process* that occurs until the system reaches a broken symmetry equilibrium stage, have developed in tandem with controlled experimental techniques in many areas of condensed matter physics (binary fluids, ferromagnets, superfluids, liquid crystals), so as to provide a solid basis for describing the dynamics of phase ordering. In cosmology, measurements of Cosmic Microwave Background anisotropies, and the formation of large scale structures in the Universe, provide signatures for phase transitions during and after *inflation*. And at the accelerator energies (Brookhaven-RHIC or CERN-LHC), phase transitions predicted by QCD could occur out of equilibrium via pion condensates.

In an instructive review on this subject, Boyanovsky and de Vega [5] describe the relevant aspects of the dynamics of symmetry breaking in many areas of physics (from condensed matter to cosmology) vis-a-vis possible experimental signatures. In condensed matter, they address the dynamics of phase ordering, emergence of condensates, and dynamical scaling. In QCD, the possibility of disoriented chiral pion condensates arising from out-of-equilibrium phase transitions is considered. And in the early Universe, the dynamics of phase ordering in phase transitions, is described, especially the emergence of condensates and scaling in Friedman-Robertson-Walker cosmologies, within a QFT framework.

## 2.4 Confinement: Supercharged Nucleus

With the failure of  $GUT$  theories to take care of the strong interaction sector  $SU(3)$  of the Standard Model, the central issue of Confinement, which has had a long history of approaches ranging from the fundamental to effective types, still remains an unsolved problem. There is a vast literature on



the subject, from Lattice QCD to various analytical methods for non-perturbative QCD. Of these, 2 novel approaches to Confinement, which are fairly self-contained, and stand out from the more conventional ones, are included in Part A, leaving the rest for Part F. The first concerns an analogy to a super-charged nucleus, based on an old work of Pomeranchuk and Smorodinsky (1940), which offers the possibility of binding a particle in a small region of space. This method was elaborated in a set of THREE "Orsay Lectures" by the late **Vladimir Gribov** [6] during 1992-94. The basic idea is that if the charge  $Z$  in a nucleus  $N_Z$  is larger than a critical value  $Z_c \approx 180$ , then this nucleus will decay to an atom of charge  $Z - 1$  and a positron:  $N_Z \rightarrow A_{Z-1} + e^+$ . If the product nucleus is unstable, the process gets repeated until the total charge of the final product is so small that further decay is impossible. Such a supercharged nucleus (a 'resonance') cannot exist freely, but only inside an atom, hence is reminiscent of a 'confined' state ! The region of stability of such a 'superbound' atomic state, (mainly due to the Pauli principle), works out as  $r_0 \ll r < 1/m$ , where  $r_0$  is the radius of the nucleus, and  $m$  the electron mass. In these three lectures, which are reproduced in this Book through the courtesy of his long term Associates Dokshitzer, Ewarz and Nyiri, Gribov [6] gives a leisurely exposition of the detailed working of this mechanism on the confinement of heavy, followed by light, quarks. These ideas have since been extended by the Dokshitzer Group in their subsequent publications hep-ph/9807224 and hep-ph/9902279, but these are outside the scope of this Book.

## 2.5 Confinement: BRST Mechanism

The second approach concerns a perspective on confinement due to Nishijima who relates its mechanism to that of an unbroken non-abelian gauge symmetry in QCD. The logic of this method which was mostly pioneered by Nishijima, may be illustrated for the case of abelian QED as follows. Quantization of the e.m. field requires "gauge-fixing", say by a covariant (Fermi) gauge. This in turn requires introduction of the indefinite (Gupta-Bleuler) metric which, for the selection of physically observable states, must be eliminated by imposing the Lorentz condition on the state vector. There are now 4 kinds of photons (2 transverse, 1 longitudinal, and 1 scalar), of which the two 'scalar' photons must have negative norms, so as to ensure manifest covariance of the quantization in the Minkowski space.

Now to project out the physical subspace, one introduces a subsidiary (Lorentz) condition (a 4-divergence of a vector field) which represents a *free, massless* field even under interactions. The photons involved in this operator (called *a*-photons) are special combinations of longitudinal and scalar photons with *zero norm*. A second (orthogonal) combination (called *b*-photons) also can be arranged to have zero norm. However the inner product of *a*- and *b*-photons is non-zero; they are 'metric partners' (somewhat akin to the 4-vectors  $n/\mu$ ,  $\tilde{n}_\mu$  defining a covariant null-plane:  $n^2 = \tilde{n}^2 = 0$ ;  $n \cdot \tilde{n} = 1$ ). A physical state is defined as one that is annihilated by applying the positive frequency part of the Lorentz condition. And since the S-matrix in QED commutes with this 4-divergence, it transforms physical states into one another, without letting them out of this subspace which now includes only *t* (transverse) and *a*-photons, but *not b*-photons. However the inner product of a physical state with one *a*-photon, with another physical state (with or without an *a*-photon), vanishes identically. Thus *a*-photons give no contribution to observable quantities, and both *a*- and *b*-photons escape detection ! This is called *confinement* of longitudinal and scalar photons in QED, a *kinematical* phenomenon !

In QCD, on the other hand, not only *a*- and *b*-gluons, but also the *t*-gluons are unobservable, giving a *dynamical* orientation to the confinement mechanism. While the basic logic and signature of confinement for non-abelian QCD remains the same as above for abelian QED, some extra ingredients of a highly technical nature are needed to bridge the gap. For not only the observable quantities now depend on the gauge parameter, but the 4-divergence of the gauge field is no longer a free field ! To eliminate the gauge-dependence of physical entities, Faddeev-Popov proposed to average the path integral over the manifold of gauge transformations, resulting in a new term in the Lagrangian (Faddeev-Popov ghost), involving a pair of anticommuting scalar fields whose violation of the Pauli theorem on spin-statistics connection requires introduction of the indefinite metric, as in QED. However, the operator analog of the Lorentz condition is more tricky in this case. It

is facilitated by a novel symmetry found by Becchi-Rouet-Stora (BRS) which was originally used for renormalizing QCD. Nishijima successfully exploited this symmetry to construct the requisite operator, and obtained a formal proof of confinement in the QCD case, as an extension of the logic employed for QED. A qualitative sketch of this proof appears in the Article by **K Nishijima and M. chaichian** [7].

### 3 Field Theory: Topological Aspects

An important sector of QFT that has come to occupy increasing importance in the last two decades, concerns its Topological aspects, as a powerful tool to probe the geometry and topology in *low* dimensions. This illustrates rather vividly the coming together of physicists and mathematicians, this time in building powerful links between quantum theory (through its path integral formulation) on the one hand, and the geometry and topology of low dimensional manifolds on the other. Indeed it appears that the properties of low dimensional manifolds can be nicely unravelled by relating them to infinite dimensional field manifolds, thus providing a powerful tool for studying these manifolds.

A unique characteristic of topological field theories is their independence of the metric of curved manifolds on which they are defined. This makes the expectation value of the energy-momentum tensor vanish. Since the only degrees of freedom are topological, there are no *local* propagating degrees of freedom. The operators are also metric independent. These features are addressed in some detail in a self-contained introductory Article by **Romesh Kaul** [8] on topological QFT regarded as a meeting ground for physicists and mathematicians.

#### 3.1 CS Theory And Jones Polynomials

Quantum  $YM$  theories in  $(2+1)D$  provide a field theoretic framework for the study of "knots and links" in a given 3-manifold, and illustrate the interplay of QFT and the topology of low dimensional manifolds. A striking result of this connection is that the famous "Jones Polynomials" of knot theory can be understood in 3D terms. This result was formally demonstrated by Edward Witten about a decade ago in a paper entitled "Quantum Field Theory And The Jones Polynomial", thus fulfilling a long-cherished goal of an exact (non-perturbative) solution of a gauge field theory, for the first time in 3 dimensions. Witten showed that the "Jones polynomial can be generalized from  $S^3$  to arbitrary 3-manifolds, giving invariants that are computable from a surgery presentation". Witten further showed that these results shed new light on 2D conformal field theory. In view of the historical importance of this pioneering work in the context of this Book theme, we reproduce (with permission from Springer-Verlag) the celebrated **Witten paper** [9] (which had appeared in *Commun.Math.Phys.***121** (1989) 351-399), in full.

#### 3.2 Anomalies In QFT

An interesting pathology of QFT which has rich topological overtones is the problem of *anomalies* which originated in the famous *ABJ* (1969) paper to resolve the problem of  $\pi^0 \Rightarrow \gamma\gamma$  decay whose hitherto standard explanation in terms of partial conservation of axial current (*PCAC*) used to fall far short of experiment. The *ABJ* paper finally resolved the issue by introducing an "anomalous" amplitude proportional to  $F_{\mu\nu}\tilde{F}_{\mu\nu}$  in the *PCAC* relation, whose interpretation brought into focus the pathology of *symmetry – breaking* at the classical level through such "anomalies" at the QFT level. Such 'violation' of gauge symmetry through 'anomalies' points to the need for their cancellation, which in turn constitutes an important constraint for physical gauge theories with *chiral* coupling to fermions. In this respect, "global chiral anomalies" play a key role in the understanding of physical effects associated with topologically non-trivial gauge-field configurations, via the celebrated Atiyah-Singer Theorem. This subject is briefly reviewed by **Haridas Banerjee** [10] in this Book.

### 3.3 Coherent States In QFT

Still another sector of QFT with topological (geometric) features, is the subject of *Coherent States* which has grown rapidly since its birth 36 years ago at the hands of Glauber and Sudarshan [R.J.Glauber, Phys.Rev.**130**, 2529 (1963); E.C.G.Sudarshan, Phys.Rev. Lett.**10**, 277 (1963)], although the basic idea dates back to the founder of Quantum Mechanics himself [Erwin Schrodinger:Naturwissenschaften, **14**, 644 (1926)] in connection with the quantum states of a harmonic oscillator, i.e., almost immediately after the birth of quantum mechanics. Coherent States have 3 main properties: coherence, overcompleteness and intrinsic geometrization, all of which play a fundamental role in QFT. These include the calculation of physical processes involving infinite number of virtual particles; the derivation of functional integrals and various effective field theories; and last not least, the exploration of the origins of topologically non-trivial gauge fields and the associated (gauge) degrees of freedom. All these topics are addressed systematically in a perspective, self-contained review by Wei-Min Zhang [11].

### 3.4 Pancharatnam-Bargmann-Berry Phase

An outstanding example of a topological aspect in quantum mechanics (which may be termed 'field theory with a finite number of degrees of freedom'), is provided by the existence of a "geometric phase" in quantum theory which had remained obscured from public view until rather recently when M.Berry (1984) drew attention to it under the term "quantum adiabatic anholonomy". Historically, however, the existence of this pathology in physics had first been noted by S.Pancharatnam (1956) in the regime of classical polarization optics, but this important work had somehow gone by default. A similar fate befell a second attempt by V.Bargmann (1964) to resurrect this idea in the context of Wigner's theorem on the representation of symmetry operators in quantum mechanics. It was only after the work of Berry that its full implications were appreciated within the physics community, but its connection with the Pancharatnam and Bargmann phases was left unattended. In an instructive Article, N Mukunda [12] describes these developments in a proper perspective by emphasizing the mutual connections among these ideas. He also describes the subsequent developments to date, by relating these phases to the presence of a complex vector space and the effect of group action among them. He then goes on to show that the geometric phase is the simplest invariant expression under certain groups of transformation acting on curves in Hilbert space.

### 3.5 Skyrmion Model for Confinement

A confinement mechanism with topological overtones is offered by the large  $N_c$  limit of QCD which has played a crucial role in unifying its premises with a solitonic, hadron-based approach that is known as the *Skyrmemodel* which was discovered by Skyrme (1961), just before quarks (1964) were born. Skyrme's novelty was to provide a model in which the fundamental fields consisted only of *pions*, wherein the nucleon was obtained as a certain classical configuration of pion fields. The apparent contradiction of making Fermi fields out of Bose fields was resolved by demanding a non-zero "winding number" for this (classical) field configuration, thus giving the "Skyrmion" the status of a topological soliton, which is a solution of a classical field equation with localized energy density.

On the face of it, the Skyrme scenario looked so different from the conventional picture of nucleons as a 'white' composite of 3 'colored' quarks bound together by their interactions with  $U(3)$  gauge fields, that a reconciliation between the two pictures appeared rather remote. It turned out however that the Skyrme model could be a plausible approximation to the orthodox QCD picture, one in which a key role is played by the large  $N_c$  limit of the latter. The logic goes roughly as follows.

Despite the increasing strength of QCD at low energies, it is plausible that the pseudoscalar mesons as  $q\bar{q}$  composites, could still interact relatively weakly with each other, thus permitting the formulation of some *effective* Lagrangian for the pions, subject of course to the correct symmetries of the underlying gauge theory, which includes a (spontaneously broken) chiral  $SU(N_f) \otimes SU(N_f)$

flavour ( $N_f$ ) symmetry that allows ‘massless’ pseudoscalars to co-exist with massive scalars. An effective Lagrangian on these lines may be obtained from “a non-linear realization of chiral symmetry”, without the explicit appearance of scalars, a structure which has an uncanny resemblance to the very Lagrangian obtained by Skyrme (1961).

How about the baryons in this QCD-motivated “chiral perturbation theory” picture? It is here that t’Hooft’s (1974) large  $N_c$  limit comes into play, with the proportionality to  $N_c$  for the baryon mass being the signal that the baryon state under study is a soliton of the effective meson theory initiated by Skyrme. In a perspective review of the Skyrme model approach, **Joseph Schechter and Herbert Weigel** [13] trace its connection with QCD in the large  $N_c$  limit, and discuss the properties of light baryons treated as solitons, within the framework of an effective Lagrangian of QCD containing only meson degrees of freedom.

## 4 Formal Methods In QFT: Selected Topics

The universal language of QFT and its powerful techniques broke fresh ground through the establishment of the equivalence of its tenets with those of Statistical Mechanics which had traditionally been developed on entirely ‘classical’ lines. In the words of A.M.Tsvelik (QFT in CMP, Camb.Univ Press 1995), this equivalence may be succinctly expressed by the following statement: “QFT of a  $D$ -dimensional system can be formulated as a statistical mechanics of a  $(D + 1)$ -dimensional system. This equivalence .... allows one to get rid of non-commuting operators and to forget about time ordering, which seem to be the characteristic properties of quantum mechanics....”. The Path Integral formulation of QFT which is the key element in dispensing with the problem of non-commuting operators in QFT, has had a crucial role in bringing about this vital correspondence of QFT with the partition function in quantum statistical mechanics (QSM). Armed with the powerful techniques of Renormalization Group Theory (RGT), this new approach has opened up a whole vista of applications to new emerging areas like critical phenomena in condensed matter physics.

### 4.1 Unified View of QFT and QSM

An important outcome of a unified view of QFT and Quantum Statistical Mechanics has been the emergence of two new areas: Euclidean Field Theory, and Finite Temperature Field Theory. Actually the origins of the former date back to the Fifties at the hands of Wick (“Wick rotation” for the Bethe-Salpeter equation) and Schwinger (as a possible direction for the evolution of QFT), wherein the transition from Minkowski to Euclidean space (via analytic continuation from real to imaginary “time”) was perceived as a means of curing many ills in QFT, such as positivity and finiteness of norms in the computation of physical quantities. In more recent times, the Euclidean formulation of QFT has led to an interesting relationship between “stochastic mechanics” (Nelson) and the Feynman-Kac formulae for Green’s functions expressed as path integrals. In a crisp Article in this Book, **R.Ramanathan** [14] provides a formulation of QFT in Euclidean space-time, to bring out the basic ideas of the Euclidean formulation, as well as the above relationship between the Nelson and Feynman-Kac formulations.

Finite Temperature Field Theory on the other hand, (in contrast to zero temperature for Euclidean QFT), provides access to a much wider class of complicated quantum mechanical systems, and addresses questions like thermal averages in QFT, symmetry restoration in theories with spontaneous symmetry breaking, and indeed the evolution of the universe at early times (from the high temperature phase). More recently, chiral symmetry-breaking phase transitions, especially the “confinement-deconfinement” phase transitions in QCD leading to quark-gluon plasmas (QGP), have acquired great interest in view of planned experiments on heavy ion collisions to detect (QGP). A few selected topics in Finite Temperature Field Theory are treated in an informative Article by **Ashoke Das** [15] in this Book.

## 4.2 Integrable Systems: Toda FT

Although most approaches to QFT have been traditionally associated with linear partial differential equations, (e.g., Schroedinger, Klein-Gordon, Dirac, Proca), *non-linear* equations, (i.e., equations where the potential term is non-linear in the field  $\phi$ ), have also been known for some time. Among the earliest non-linear wave equations known in physics are the *Liouville* and *Sine - Gordon* equations. The Liouville equation in 2D arose in the context of a search for a manifold with constant curvature, something like covering the surface with a fishing net whose arc length is constant (knots do not move!), while the 'threads' in the net correspond to a local coordinate system on the surface. The "field"  $\phi$  in the Liouville equation is the phase space density  $\rho$  satisfying the equation  $\partial_x \partial_y \rho = \exp \rho$ , where  $x, y$  are the local orthogonal coordinates. The Sine-Gordon (SG) equation has a similar structure, with  $\exp$  replaced by  $\sin$  on the RHS. Variants of these equations, e.g., adding a 'mass' term  $m^2 \phi$  on the LHS, and/or the hyperbolic replacement of  $\sin$  by  $\sinh$ , etc., give rise to several more varieties of similar types. A third type of non-linear equation which has received much attention, is the so-called *KdV* equation  $u_t - 6uu_x + u_{xxx} = 0$ , with interesting properties like an infinite number of conservation laws. The corresponding conserved quantities can be used as Hamiltonians for an integrable system (*KdV* hierarchy). A striking feature of such non-linear equations is an infinite number of conserved quantities, which imply that the solutions of these systems must be infinitely restricted. This results in such solutions being quite stable structures (*solitons*) which retain their shapes even after collisions.

An interesting class of coupled non-linear equations was introduced by M.Toda (1967) to describe a 1D crystal with non-linear coupling between nearest neighbour atoms. These (lattice) models also admit *soliton* solutions which reduce to the *KdV* equation in the continuum limit. At the 'field' level, such models (with exponential 'potentials') simulate a general class of non-linear equations-called Toda Field Theory which include the Liouville and Sine-Gordon equations as special cases. For the solution of these equations, a general method of "inverse scattering" was proposed by Gelfand-Levitan. The logic of this method is to convert, via a suitable transformation, the original non-linear equation to an equivalent *linear* equation, and study the evolution of the latter, more or less according to standard methods already developed for them (including group-theoretic, Lie-algebraic, etc methods). The inverse scattering method paved the way to connections with other known models of QFT, such as conformally invariant FT and the Hamiltonian reduction of Wess-Zumino-Witten model. Similarly the *KdV* equation is related to the 4D Yang-mills theories, thus providing a connection of the latter with 2D integrable models. In an instructive, self-contained article on this subject, **Bani Sodermark** [16] gives a perspective view of integrable systems with special reference to the Toda Lattice hierarchy, and reveals the connections of such non-linear field theories with other sectors of QFT.

## 4.3 Light-Front Dynamics

Dirac laid the foundations of QFT, not only through his famous Equation, but at least with 2 more seminal contributions within a year's gap from each other: a) light-front (LF) quantization [Rev.Mod.Phys.**21**, 392 (1949)]; b) constrained dynamics [Can.J.Math.**2**, 129 (1950)]. In the former, he suggested that a relativistically invariant Hamiltonian theory can be based on different classes of initial surfaces: instant form ( $x_0 = \text{const}$ ); light-front (LF) form ( $x_0 + x_3 = 0$ ); hyperboloid form ( $x^2 + a^2 < 0$ ). The structure of the theory is strongly dependent on these 3 surface forms. In particular, the "LF form" remains invariant under 7 generators of the Poincare' group, while the other two are invariant only under 6 of them. Thus the LF form has the maximum number (7) of "kinematical" generators (their representations are independent of the dynamics of the system), leaving only 3 "hamiltonians" for the dynamics.

Dirac's LF dynamics got a boost after Weinberg's discovery of the  $P_z = \text{inf}$  frame which greatly simplified the structure of current algebra. The Bjorken scaling in deep inelastic scattering, supported by Feynman's parton picture, brought out the equivalence of LF dynamics with the  $P_z = \text{inf}$  frame. The LF language was developed systematically within the QFT framework by Kogut-Soper (1970), Leutwyler-Stern (1978), Srivastava (1998) and others. The time ordering in

LF-QFT is in the variable  $\tau = x_0 + x_3$ , instead of  $t = x_0$  in the instant form. And despite certain technicalities, the LF dynamics often turns out to be simpler and more transparent than the instant form, without giving up on the net physical content. This is borne out from comparative studies: of spontaneous symmetry breaking on the LF; of degenerate vacuum in certain  $(1+1)D$  QFT which are exactly soluble and renormalizable (e.g., the Schwinger model and its chiral version); of chiral boson theories; and of QCD in covariant gauges. Indeed, the LF quantization of QCD in the Hamiltonian form bids fair to be a viable alternative to the lattice gauge theory for calculating non-perturbative quantities. Removal of constraints by the Dirac method gives fewer independent dynamical variables in the LF formalism than in the instant form; for this reason, LF variables have found applications even in String and  $M$ -theories. In an instructive self-contained review (with a rich collection of references), **Prem Srivastava**[17] gives a detailed review of most of these topics in a leisurely and systematic manner, and leads the interested reader all the way to the frontier with several new results.

#### 4.3.1 2D Field Theory

$2D$  models in QFT have also been of great interest in the contemporary literature. Such theories reveal some remarkable features, such as fermion-boson equivalence, which facilitates the solution of fermion-FT in terms of its bosonized version. This concept of bosonization in turn has been useful in the understanding of  $4D$  phenomena that can be described by an effective  $2D$  FT, such as the demonstration of quark confinement in exactly soluble  $2D$  models [Casher-Kogut-Susskind (1973)]. Another important discovery in  $2D$  FT concerns an "anomaly-generated" mass [Jackiw-Rajaraman (1985)] for the gauge boson in the Chiral Schwinger model. (This mechanism may be contrasted to the standard Higgs mechanism for generating the vector boson mass via spontaneous symmetry breaking). The "anomaly" here stands for the loss of the conservation property due to quantum corrections involved in the quantization of the gauge theory. This disease in turn needs Dirac's second weapon for cure: Constrained dynamics. In a short perspective article in this Book, **Dayashankar Kulshreshtha** [18] reviews the constrained dynamics and local gauge invariance of several  $2D$  FT models, in both Instant and LF forms, and in so doing, brings out the detailed working of the BRST formalism as applied to such  $2D$  models.

### 4.4 Constrained Dynamics

To recall the essential elements of a constrained dynamical system, which includes most systems of physical interest (e.g., QED, QCD, Electroweak and Gravity theories), it is characterized by an *over-determined* set of coordinates. These are best kept track of within a Hamiltonian formulation, which has a natural place for all the coordinates (canonical and redundant), so that the complete set of constraints emerges easily. The nature of these constraints in turn is determined by the structure of the matrix of Poisson brackets (PB) of the constraints of the theory, which also carries the signature of whether or not the underlying theory is gauge invariant (GI). Thus if this PB matrix is singular, then the set of constraints is *firstclass*, and the theory is GI. On the other hand, if this matrix is non-singular, then the set of constraints of the theory is *secondclass*, and the theory is non-GI. (Indeed this is often taken as a criterion for distinguishing a GI from a non-GI system [18]). These GI systems are then quantized under some appropriate gauge choices, or "gauge fixing" (GF). Now in the usual Hamiltonian formulations of a GI theory under some GF's, one necessarily destroys the gauge invariance, since the GF corresponds to converting the first class constraints to second class constraints. To quantize a GI theory by maintaining gauge invariance despite GF, one needs the more general BRST (1974) formulation, wherein the theory is rewritten as a quantum system with generalized GI, called BRST invariance. This in turn requires enlarging the Hilbert space, and replacing the gauge transformation by a BRST transformation which involves the introduction of (anti-commuting) Faddeev-Popov *ghostfields*. This amounts to embedding the GI system into a BRST invariant system (but isomorphic to the former), whose unitarity is guaranteed by the conservation and nilpotency of the BRST charge.

Thus the Dirac[Can.J.Math.2, 129 (1950)]-Bergmann [Phys.Rev.83, 1018 (1951)] theory of



Constraints lies at the root of (Hamiltonian) description of interactions in QFT based on Action principles which, due to the requirements of Lorentz, local gauge, (and/or diffeomorphism) invariances, must employ singular Lagrangians. This is generally adequate for the study of simple gauge theories (controlled by some Lie groups acting on some internal space in Minkowski space-time), via the covariant approach based on BRST symmetry which, at least for infinitesimal gauge transformations, allows a regularization and renormalization of the relevant theories within the local QFT framework. On the other hand, the gauge freedom of theories that are invariant under diffeomorphism groups of the underlying space-time (e.g., in general relativity or string theory) is encumbered by the arbitrariness for the observer in the "definitory properties" of space-time and/or the measuring apparatus; [see **L. Lusanna**-this Book]. Such ambiguities affect bigger issues like: the understanding of *finite* gauge transformations; the Gribov ambiguity in the choice of function space for the fields; proper definition of relativistic bound states vis-a-vis quark confinement; and last not least the conceptual and practical problems posed by gravity. These require a fresh look at the foundations of QFT to know if we: i) understand the physical degrees of freedom hidden behind gauge and/or general covariance; ii) can meaningfully reformulate the physics (both classical and quantum) in terms of them. Logically this would amount to abandoning local QFT for non-perturbative interactions, and a reformulation of relativistic theories to allow natural coupling to Gravity. These and allied issues are addressed in a state of the art review by **Luca Lusanna** [19], aimed at a unified reformulation of the 4 basic interactions in terms of Dirac-Bergmann observables, with emphasis on the open problems mathematical, physical and interpretational.

## 5 Extension Of QFT Frontiers

A long term ambition of QFT has been the dream of unification of all the gauge fields with the Gravitation Field whose quantization has all along posed a big challenge in its own right. [A major difficulty in the way of unification of this sector with the other three, as was once succinctly put by Abdus Salam, lies in the "spin mismatch" of their respective fields (vector vs tensor), which would militate against a common strategy]. Nevertheless such a unification was to come about from an entirely new paradigm which envisaged extension of the original tenets of Field Theory based on a point particle description to one with *Strings*. In this Section we offer a panoramic view of some major theoretical developments from seemingly unrelated angles, which, apart from their impact on Physics in their own right, have provided some key ingredients converging towards the emergence of modern *StringTheory*. These developments which may be termed *Supersymmetry* (SUSY), *ConformalFieldTheory* (CFT), and *Duality*, are outlined next.

### 5.1 SUSY In Field Theory

In its march towards Unification, Field Theory has continued to break new ground in several directions. An important step in Unification was marked by the discovery of Supersymmetry (*SUSY*), introduced in the early seventies by a galaxy of authors in the context of 2D QFT (Gervais-Sakata) as well as in 4D QFT (Golfand, Likhtman, Akulov, Volkov, Wess and Zumino), for a unified understanding of the two known forms of matter—bosons (integral spins) and fermions (half integral spins)—hitherto regarded as two distinct field types, with commuting and anticommuting properties respectively. The new symmetry between bosons and fermions may be incorporated within the definition of a single "Superfield", with transformations inter-relating the two constituents, so that *SUSY* becomes a part of space-time symmetry implied by relativistic invariance. The Gauge principle too admits of a corresponding extension to unify both these sectors.

What are the motivations for such a lavish extension of space-time symmetry ? Apart from its aesthetic appeal, there are some theoretical considerations of a more concrete nature which are dwelt on in this Book through two complementary reviews of *SUSY* in Field Theory, (with special reference to Particle Physics), by two leading experts in the field: **Rabi Mohapatra** [20] and **Norisuke Sakai** [21] respectively. According to Sakai [21], the most important motivation for *SUSY* is the *Gaugehierarchyproblem* showing up via the vastly different mass scales of the

electroweak ( $M_W$ ) vs the "GUT-theoretic" ( $M_G$ ):  $M_W^2/M_G^2 \approx 10^{-28}$ . A similar gap exists between the "GUT" vs Planck (gravity) mass scales:  $M_G^2/M_{Pl}^2 \approx 10^{-6}$ .

To account for this phenomenon, it is necessary to invoke a suitable *Symmetry* reason which may be precisely formulated by the so-called "naturalness" hypothesis (t'Hooft 1979) which demands that a system acquires a higher symmetry as a certain (small) parameter goes to zero, e.g., chiral symmetry occurs when a (small) fermion mass goes to zero; or a local gauge symmetry corresponds to the vanishing of a vector boson mass. Now the mass scale  $M_W$  of weak bosons arises from the vacuum expectation value  $\langle \phi \rangle_0 \equiv v \neq 0$ , related to the mass  $M_H$  of the Higgs scalar field  $\phi$ . So to regard the gauge hierarchy problem as the result of some symmetry breaking, we must give a *Symmetry* reason to make the Higgs scalar mass vanishingly small. Classically a vanishing scalar mass corresponds to a symmetry called scale invariance, which however cannot be maintained quantum mechanically. In a perspective review on this subject, **Norisuke Sakai** [21] argues for "Supersymmetry" between the Higgs scalar and a spinor partner as a good option: Chiral symmetry gives zero mass to the latter, while *SUSY* makes the former massless (through a cancellation of the respective contributions to the self-energy loops).

In a complementary perspective review on the same subject, **Rabi Mohapatra** [20] stresses the versatility of *SUSY* as a tool for understanding many unsolved problems of physics: a) improvement in the singularity structure of local fields for understanding the disparate scales of Nature (e.g., Electroweak vs Gravity); b) possibility of unifying Gravity with the other forces by making *SUSY* local instead of global; c) prospects of understanding *non-perturbative* properties of field theories, hitherto considered 'impossible' in non-*SUSY* form.

As to the manifestations of *SUSY* in a real world, this "Bose-Fermi" symmetry is supposed to be badly broken, so that any search for superpartners (bosons vs 'bosinos'; fermions vs 's-fermions') has so far yielded zero dividends. On the other hand the formulation of Supersymmetry in non-relativistic quantum mechanics is relatively free from constraints. Indeed, since Schroedinger (1940) noticed the existence of well-defined "supersymmetric partners" for the energy levels of a given quantum mechanical system, many applications to such systems (including nuclear and condensed matter physics), have kept pace with the rapid strides of *SUSY* in field theory in recent years. Indeed, the existence of *SUSY* partners in the energy levels of (appropriately chosen) even vs odd nuclei have been systematically established by group theoretic methods (interacting boson models, etc). Similarly, in solid state physics, an interesting correspondence has been observed between the critical behaviour of a 'spin system' in random magnetic fields in  $d$  dimensions, and that of the spin system without the random magnetic field in  $d-2$  dimensions. This "dimensional reduction" may be traced to an underlying *SUSY* for the spin system in random magnetic fields (see **N Sakai**-this Book).

In the absence of discovery of *SUSY* partners in Field Theory, the benefits from *SUSY* have so far been purely theoretical, varying from reduction of the degrees of divergence arising from various loop integrals in standard field theory (by at least two orders), to a heavy reduction in the number of dimensions (from 26 to 10) needed for self-consistency in a string theoretic formulation. The Articles by Mohapatra [20] and Sakai [21] between them provide quite a complementary description of the *SUSY* formalism in QFT, together with an glimpse of the recent developments. And apart from its applications in particle physics, this formalism also serves as a background to the vast field of supersymmetric string theory.

## 5.2 CFT

An independent insight into the origin of String Theory comes from the role of *Conformal Field Theory* (CFT), viz., conformally invariant QFT in 2D(imensions), not only as a vital ingredient of its anatomy, but also with firm hold on other disciplines like condensed matter physics. The CFT route to the evolution of String Theory is sketched in this Book as part of a bigger (historical) survey by **Werner Nahm** [22], tracing a whole sequence of developments in QFT right from its (Dirac) beginning, and encompassing in the process several other areas of physics on which CFT has had a decisive impact. In this saga, the interplay of physical intuition and mathematical rigour has brought together the practitioners of these respective disciplines, though not necessarily

working in tandem. On the one hand, the beauty and transparency of CFT have made for a rich variety of intellectual exercises in abstract mathematics (with new emerging areas like automorphic groups, Kähler-Einstein metrics, etc), and on the other, facilitated the study of intensely practical physical systems such as continuous phase transitions in condensed matter physics.

The impact of CFT on string theory has had its origin in several theoretical developments: the Thirring model in 2D; Skyrme's idea of the equivalence between Fermions and Bosons; Coleman's equivalence theorem on the Thirring Model versus the Sine-Gordon equation (despite their apparent dissimilarity); and the role of conformal invariance in the structure of Wilson's Renormalization Group equations. To recall the essentials of Conformal invariance, this symmetry is satisfied in the absence of any 'scale' dimension. Examples are Maxwell's Equations in free space; Dirac equation for massless fermions which satisfy conformal invariance. The 2D Thirring model, which may be regarded as a basic ingredient of string theory, also has this property due to absence of a scale dimension. Using this mathematical picture, the string may be regarded as a 1D object in space spanning a world sheet (a Riemann surface) embedded in 2D space-time, where a point on the string is represented by  $X^\mu(\sigma, \tau)$ ;  $\sigma, \tau$  being the 2 world sheet coordinates.

The impact of CFT has been no less impressive in the domain of condensed matter physics (CMP) where there exist a rich class of QFT's exhibiting the structure of conformally invariant fields, such as in 2D surface coatings. Thus at a critical temperature ( $T_c$ ), the long range fluctuations of arbitrary scales make irrelevant the details of molecular structure, and the theory approaches a *continuum limit*, with no visible scale dimension to keep track of. Indeed in this limit, the correlation functions behave like the Euclidean  $n$ -point functions of standard QFT with conformal invariance properties. Nahm [22] discusses an interesting correspondence between the Ising model in CMP and Thirring model in QFT. The equation satisfied by the spin waves of the Ising model is formally identical to the 2D Dirac equation for massless fermions. Indeed condensed matter physics provides a more stable and economical background for testing these ideas than the expensive HEP laboratories !

### 5.3 String Theory Via Duality

Perhaps the most startling "revolution" in Physics to date which had its origin in QFT, has been the String Theory, and its successive "Avatars" (incarnations), aimed at unifying all the forces of Nature (from orthodox gauge theories of strong, e.m. and weak interactions, all the way to gravity). An orthodox route to its evolution may be attributed to the strong interaction problem in QFT, which has had wide ramifications from vastly different angles, each providing an independent insight into its mysteries. A very promising approach to strong interactions came from the *Duality* Principle which has had a long history (perhaps traceable to the Bootstrap hypothesis), based on the equivalence of the direct channel (resonances) and crossed channel Regge poles with a universal slope  $\alpha' \approx 1\text{GeV}^{-2}$ . An explicit realization of this idea was achieved via the Veneziano representation for 4-point amplitudes satisfying the requirements of duality and crossing symmetry, which was soon generalized to  $N$ -point amplitudes satisfying the same properties. Through a path integral representation of such amplitudes, Nambu, Nielsen and Susskind recognized that these amplitudes describe a 1D (string-like) object moving in space, with the inverse of the universal Regge slope identified as the "string" tension  $T$ . The "string" interpretation was further reinforced by a subsequent representation due to Virasoro, with very similar properties. And its promise of relevance to particle physics (despite stiff competition from QCD!) got a boost from the Scherk-Schwarz (1974) observation that such a "string theory" could serve as a candidate for incorporating gravity in its ambit, on the ground that the massless spin-2 particle appears naturally in the closed string spectrum. To that end the string tension  $T$  needed to be increased by 19 orders of magnitude (up to the Planck scale!) to qualify for a viable theory of gravity. The conceptual gap was finally bridged by the seminal work of Green-Schwarz (1984) who succeeded in constructing a consistent 10D super Yang-Mills theory coupled to supergravity which is free from anomalies only for certain gauge groups ( $SO(32)$  or  $E_8 \times E_8$ ). This work, perhaps for the first time, showed real prospects for unifying the fundamental forces.

The String Theory has grown by leaps and bounds during the past decade, and its vast ram-

ifications have grown to such formidable literature over the past decade, that a minimal justice to it would itself require several volumes of review. Nevertheless, after a short overview by the Master, **John Schwarz** [23] of the subject, a panoramic account of the major developments in this exciting field (together with an exhaustive set of references) is given in a perspective Article by **Jnanadeva Maharana** [24]. Schwarz [23] views the different superstring theories (and an extension called  $M$ -theory) as different facets of a unique underlying theory going beyond ordinary QFT's. However, recent duality conjectures suggest that a more complete definition of these theories may come from the large  $N$  limits of suitably chosen  $U(N)$  gauge theories; (see **L Bonora** [25] below). The Maharana Article [24] leads the interested reader through several stages of its development, from i) perturbative aspects of  $ST$ ; successively through ii) *DualitySymmetries* as a characteristic of String Theories (ST); iii)  $M$ -theory as a unified view of the *five* perturbatively consistent  $ST$ 's; iv) microscopic understanding of Black Holes, and so on, all the way to the frontiers of the field.

Attempting to cover the later stages of development in this rapidly growing field, **Loriano Bonora** [25] reviews some advances in the study of the relation between Yang-Mills ( $YM$ ) theory and strings, based on the classical  $YM$ -theory solutions (*Riemannian instantons*) which are 2D solutions describing Riemann surfaces in the strong coupling limit. Strictly, such relations historically date back all the way to 't Hooft ('74) through his famous  $1/N_c$  expansion for large  $N_c$ , wherein the dominant Feynman amplitudes correspond to the 2D Riemann surfaces. This 'natural connection' with strings was subsequently upgraded to a concrete shape via studies of 2D QCD (for string-like properties), which was further generalized to a connection between conformal super- $YM$  and super-string theory of type  $IIB$ , in the large  $N_c$  limit. The **Bonora** Article reveals, among other things, a direct link between String Theory and non-abelian  $YM$  theory, through the emergence in the latter of classical solutions modelled over Riemann surfaces, leading to a "string" interpretation. Historically, this came about only after the proposal of the *MatrixTheory*, which in the large  $N_c$  limit converges to the (non-perturbative)  $M$ -Theory.

## 6 CS Field Theory And Condensed Matter Physics

While the dominant concern in Field Theory has been in the traditional domain of particle physics, its powerful language and techniques have found profitable employment over a much wider domain, which comprises topics in Condensed Matter Physics, and newly emerging fields like Quantum Hall Effect, fractional statistics and *Anyons*. These phenomena lend themselves to QFT treatment in  $(2+1)$  dimensions, where the celebrated "Chern-Simon" ( $CS$ ) term plays a key role (see also Part B on Topological Field Theories).

What are the special features of QFT in  $(2+1)D$ , and what specific role does the  $CS$  term play in this reduced space-time continuum? Perhaps the most striking feature is the appearance of *fractionalstatistics*! For, whereas in 3 (or higher) space dimensions, all particles must either be bosons (integral spin) or fermions (half-integral spin), in 2 space dimensions, the particles can have any fractional spin/statistics with impunity! Such particles are called *Anyons*. Now since the usual spin-statistics relation follows from the premises of the standard 4D relativistic QFT, it is natural to ask if *Anyons* can be understood from the corresponding 3D QFT. The question goes beyond mere academic interest since lower dimensions can be effectively realized in the physical world through the "freezing" out of certain degrees of freedom, (e.g., in a strongly confined potential, or at low enough temperatures), so that these 'quasi-particles' may well exhibit anyon-like properties. And indeed experiments on Quantum Hall Effect (QHE) have revealed the existence of fractionally charged excitations (thus implying anyons).

A critical discussion on the question of anyons and fractional statistics in  $(2+1)$  dimensions, with particular reference to the role of the Chern-Simons ( $CS$ ) term in 3D QFT, is given by **Avinash Khare** [26] in a perspective Article on the subject in this Book. To that end, Khare clarifies the definition of "quantum statistics" which relates to the "phase" picked up by a wave function when two identical particles are *adiabaticallyexchanged*, as distinct from the usual definition of permutation symmetry for two identical particles. [While both definitions coincide for 3 and higher

dimensions, they differ in 2 dimensions]. He then discusses in detail the main properties of the  $CS$  term, especially its role as a gauge field mass term, in whose presence anyons can appear in one of two different ways: i) as a soliton of the corresponding QFT ; or ii) as fundamental quanta carrying fractional statistics. So far, the state of the art is based on non-relativistic QFT, wherein the  $CS$  term provides an effective cushion against a non-local formulation of anyon fields, thus facilitating a 'local' formulation. However a full-fledged relativistic QFT formulation is not yet feasible.

Perhaps the most tangible success from  $CS$  fields so far is a natural understanding of the Quantum Hall (QH) Effect. A state-of-the-art review by **R Rajaraman** [27] puts this topical subject in perspective. We summarize some essential features of a QH system, from his own account. A QH system which is defined as "quasi 2D layers of electrons trapped in the interface of semi-conductors, at very high magnetic fields and very low temperatures, has revealed many remarkable features ". Particularly interesting is the presence of certain states characterized by the so-called "filling fractions" ( $\nu$ ) which are either integers, or certain *odd denominator* fractions;  $\nu = hc\bar{\rho}/eB$ , where  $\bar{\rho}$  is the mean electron density, and  $B$  the applied field. The special states corresponding to these  $\nu$ -values show extremely flat plateaus in Hall conductivity which (in units of  $e^2/h$ ) are exactly equal these values to within an accuracy of 1 in  $10^7$  ! These features are very universal inasmuch as the details of the material seem irrelevant. It was earlier recognized that the electrons in these QH states form an incompressible fluid, described by "Laughlin wave functions" (which are reminiscent of Jastrow-type correlations in nuclear wave functions). A more analytical study of these empirical functions suggested a Landau-Ginsberg type scenario for the QHE in terms of an order parameter field (subsequently to be identified with a Chern-Simons field), thus formally bringing this subject within a 3D QFT network.

The analogy of the order parameter field in QHE to that obtaining in superconductivity of the Landau-Ginsberg description, is of course not a literal one since there are no bosonic Cooper pairs in QHE. Indeed in this 3D QFT scenario, the "anyons" (Chern Simons fields) have an intermediate status between bosons and fermions. However for the special case when the anyon angle is an odd multiple of  $\pi$ , a composite of the electron with an *odd* number of flux tubes, effectively amounts to constructing a "bosonic" analogue of Cooper pairs from out of "fermions" which now provides the desired order-parameter (CS) field operating in the plateau of the QH system. Rajaraman reviews a formal QFT procedure for constructing such CS gauge fields, as well as the formulation of their dynamics at the 3D QFT level. As to the connection of the CS gauge fields at the first quantized level, these are of course expressible in terms of the "phase angles" involved in the exchange of electrons in an  $N$ - electron wave function in 2D (see also Khare [26] in this Book).

## 7 QCD-Motivated Strategies For Strong Interactions

Turning now to the strong interaction problem in the standard field theoretic picture, its prime candidate, QCD, has since its birth been beset with problems of reliable calculational techniques to deliver results. An introductory overview of several approaches [symmetries, effective Lagrangians and Wilson expansions] to deduce hadron properties from QCD is sketched in the Article by **Olivier Pene** [28], aimed at establishing a link between perturbative and non-perturbative QCD via lattice methods. We now go into more specific details of a few principal QCD-based methods.

### 7.1 QCD Sum Rules

To recall the main signatures of the prime candidate, QCD, which it shares with any non-abelian gauge theory, are expressed by a two-fold pattern: i) decreasing coupling strength at shorter distances (Asymptotic Freedom); and ii) increasing coupling strength at longer distances (*confinement*). The former is fairly well understood, and provides a perturbative basis for calculating QCD effects in high energy processes. In particular, the powerful method of "QCD Sum Rules", based on Wilson's Operator Product Expansion (*OPE*), was developed by Shifman-Vansteijn-Zakharov for the study of non-perturbative QCD in a large variety of applications from hadronic masses (with two-point functions), coupling constants, form factors (with three-point

functions), and reactions (four point functions). The basic philosophy is one of a *duality* between two ways of representing a correlator: i) *OPE* with various "twist" terms (vacuum condensates, treated as free parameters of the theory) representing successive non-perturbative corrections to an otherwise perturbative expansion; ii) a dispersion formula saturated by certain low-lying hadron resonances. Equating the two amounts to evaluating hadronic parameters in terms of the quark-level condensates. Despite certain conceptual problems of "microscopic causality" encountered in the "matching" of two sides of the equation, this method (QCD-SR) has proved very popular among a wide class of high energy phenomenologists, and has been continually refined over the years. A leisurely review of the state of the QCD-SR art on the quark structure of hadrons, as well as its working on the problem of hadrons in nuclear matter (at finite temperature) is given by **Leonard Kisslinger** [29] in this Book.

## 7.2 Non-Perturbative Methods With QCD Features

The state of the art in this field is so diffuse that a more organized exposition is needed for such methods. To that end the attempts at addressing the strong interaction problem in QCD may be divided into two broad categories: i) soluble models designed to shed light on its general features through exact calculations; and ii) effective Lagrangian methods for 4-fermion interactions, somewhat reminiscent of the Bethe "Second Principle" Theory of effective nucleon-nucleon interactions of the Fifties. Srivastava [17], as well as Kulshreshtha [18], in Part C of this Book, have already provided a flavour of the results to be expected from type (i) theories, using the method of LF-QFT.

Type (ii) which deals with more realistic situations, albeit at the cost of some phenomenology, has a much wider literature to choose from. To do a semblance of justice to this field, this Book includes *two* articles of this type, reviewing the methodology and working of such QFT-based approaches. The first one, by **Vladimir Karmanov** [30], gives an in-depth review of covariant light-front (LF) dynamics, with applications to field theory and relativistic wave functions. The formalism is effectively 3D in content, which can be obtained by projecting the (4D) Bethe-Salpeter amplitudes on the light-front plane, and although a reversal of steps is not possible to reconstruct the 4D BS amplitude, the LF formalism still represents a powerful alternative for solving QFT problems. Karmanov [30] also discusses some typical applications.

### 7.2.1 Markov-Yukawa Transversality Principle

The second article by **Asoke Mitra** [31] offers a comparative view of the state of the art in several QFT approaches based on effective 4-fermion interactions (including QCD features), of both 3D and 4D types (Tamm-Dancoff, Bethe-Salpeter, Salpeter, quasi potentials, light-front). In this context, attention is focussed on an important but somewhat less known principle called "Markov-Yukawa Transversality" (*MYTP*) which decrees that the interaction between the two (quark) constituents be *transverse* to the composite (hadron) 4-momentum, by virtue of which the BSE kernel has an effective (albeit covariant) 3D support. As a result of this "Covariant Instantaneity" the starting 4D BSE is exactly reducible to a 3D form, and conversely the steps can be reversed so as to allow an *exactreconstruction* of the original 4D BSE in terms of 3D ingredients ! Thus *MYTP* allows an exact interlinkage between the 3D and 4D BSE forms, so that both forms can be used interchangeably, unlike most other approaches in the literature which employ either a 4D or a 3D form of the BS dynamics, but not both simultaneously.

It might be of some historical interest to note that the Salpeter equation has a 3D structure stemming from its (instantaneous) kernel with a 3D support, and therefore its original 4D form can be recovered a la *MYTP* by reversing the steps, but this possibility had never been explored. This gap is now filled by *MYTP* which provides a formally covariant basis for the instantaneous approximation. The same principle (*MYTP*) can also be generalized from covariant instantaneity to the covariant light-front.

A fall-out of the 3D-4D interlinkage provided by *MYTP* is that it gives a *two-tier* description: the 3D form for the hadron spectra which are  $O(3)$ -like; and the 4D form to address the transition



amplitudes as 4D loop integrals using standard (4D) Feynman rules. This Principle can be easily incorporated in the usual framework of coupled Bethe-Salpeter and Schwinger-Dyson equations (BSE-SDE) stemming from a (chirally invariant) 4-fermion Lagrangian with current quarks interacting via the full gluon propagator, so that the quark mass is acquired via the NJL-mechanism. And the generalization from covariant instantaneity to the covariant light-front helps remove certain problems of Lorentz mismatch of vertex functions that arise in a 4D loop integral under the covariant instantaneity ansatz. These and other details are reviewed in the article by Mitra [31] which also stresses a parallelism of treatment of  $q\bar{q}$  and  $qqq$  systems.

### 7.3 The Harmonic Oscillator: A Powerful Bridge In QFT

No amount of literature on the impact of QFT in Physics would be complete without an exposure of the role of the Harmonic Oscillator (HO) in shaping Quantum Theory, as an integral part of this Book theme. It was therefore a matter of great satisfaction when **Marcos Moshinsky** [32], who may be regarded as the "Father of the Harmonic Oscillator in Physics", agreed to contribute a perspective article on the HO theme. The only obstacle against a regular format for his Article was that he had only recently written a comprehensive book on the subject [M. Moshinsky and Yu.F.Smirnov, *The Harmonic Oscillator In Modern Physics*, (Harwood Academic Press, the Netherlands, 1996)]. Nevertheless in his Article, he has provided a comprehensive list of contents of his HO-book, which already offers a glimpse of the depth and range of physical problems (from the simplest quantum mechanical ones to the  $n$ -body Relativistic Oscillator) that are amenable to the amazing powers of HO techniques in tandem with the standard methods of Group Theory. In addition he has reviewed some recent work of his on relativistic particles of arbitrary spin in a *confining* HO potential, with applications to Spectroscopy.

## 8 Conclusion: Foundations Of Quantum Theory

The Book concludes with an Article by **Dipankar Home**[33] on the foundations of quantum mechanics (the predecessor of QFT). It is well known that the Founding Fathers of quantum theory (Planck, Einstein, and Schroedinger) held deep reservations about the adequacy of quantum theory as a complete description of Nature. Since Quantum Field Theory is the application of the same quantum principles to systems with continuous degrees of freedom, it is open to the same questions. Indeed the very premises of quantum theory are now increasingly being scrutinized by relating them to precise experimental studies. Home's Article [33] concentrates on two main issues: i) the measurement problem in quantum theory; and ii) quantum non-locality, both being areas of active research which are yielding new and unexpected results. He concludes with a quotation from John Bell: "It seems to me possible that the continuing anxiety about what quantum mechanics means or entails will lead to still more tricky experiments which will eventually find some soft spot." Translated to the QFT level, this looks like an appropriate conclusion for this Book as well.

## References

- [1] D.V.Shirkov: Evolution Of The Bogoliubov Renormalization Group
- [2] S. Szpigel and R.J.Perry: The Similarity Renormalization Group
- [3] V.Novikov: Quantum Field Theory And The Standard Model - Bird's Eye View
- [4] P.K.Kabir: Broken Reflection Symmetries
- [5] D.Boyanovsky and H.J.de Vega: Dynamics Of Symmetry Breaking Out Of Equilibrium - From Condensed Matter To QCD And The Early Universe
- [6] V.N.Gribov (Orsay Lectures): (I) hep-ph/9403218; (II) hep-ph/9404332 ; (III) hep-ph/9905285

- [7] K.Nishijima and M.Chaichian: An Essay On Color Confinement
- [8] R.Kaul: Topological Quantum Field Theories - A Meeting Ground For Physicists And Mathematicians
- [9] E.Witten: Quantum Field Theory And The Jones Polynomial
- [10] H.Banerjee: Chiral Anomalies In Field Theories
- [11] Wei-Min Zhang: Coherent States In Field Theory
- [12] N.Mukunda: Pancharatnam, Bargmann And Berry Phases - A Retrospective
- [13] J.Schechter and H.Weigel: The Skyrme Model For Baryons
- [14] R.Ramanathan: Euclidean Methods In Quantum Field Theory
- [15] Ashoke Das: Topics In Finite Temperature Field Theory
- [16] B.M.Sodermark: Integrable Models And The Toda Lattice Hierarchy
- [17] P.P.Srivastava: Perspectives Of Light-Front Quantized Field Theory - Some New Results
- [18] D.S.Kulshreshtha: Gauge Symmetry In Chiral Electrodynamics
- [19] L.Lusanna: Towards A Unified Description Of The Four Interactions In Terms Of Dirac-Bergmann Observables
- [20] R.N.Mohapatra: Supersymmetry And Particle Physics
- [21] N.Sakai: Supersymmetry In Field Theory
- [22] W.Nahm: Conformal Field Theory: A Bridge Over Troubled Waters
- [23] J.H.Schwarz: Superstring Theory - An Overview
- [24] J.Maharana: Recent Developments In String Theory
- [25] L.Bonora: Yang-Mills Theory And Matrix String Theory
- [26] Avinash Khare: Fractional Statistics And Chern-Simons Field Theory In  $2 + 1$  Dimensions
- [27] R.Rajaraman: Chern Simons Field And Composite Bosons In The Quantum Hall System
- [28] O.Pene: Hadrons From QCD - Achievements And Prospects
- [29] L.S.Kisslinger: QCD Sum Rules In Hadronic And Nuclear Physics
- [30] V.A.Karmanov: Light-Front Dynamics
- [31] A.N.Mitra: 3D-4D Interlinkage Of Bethe-Salpeter Amplitudes - A Unified View Of  $Q\bar{Q}$  And  $QQQ$  Dynamics
- [32] M.Moshinsky: The Harmonic Oscillator In Quantum Theory - A Powerful Bridge In Physics
- [33] D.Home: Modern Perspectives On Foundations Of Quantum Mechanics





# Part A : Basic Structure Of QFT

1. Evolution Of The Bogoliubov Renormalization Group by D.V.Shirkov
2. The Similarity Renormalization Group by S.Szpigel and R.J.Perry
3. Quantum Field Theory And The Standard Model - Bird's Eye View by V.Novikov
4. Broken Reflection Symmetries by P.K.Kabir
5. Dynamics Of Symmetry Breaking Out Of Equilibrium - From Condensed Matter To QCD And The Early Universe by D.Boyanovsky and H.J.de Vega
6. Orsay Lectures On Confinement by the Late Vladimir.N.Gribov: (I) hep-ph/9403218; (II) hep-ph/9404332 ; (III) hep-ph/9905285 (Courtesy of Y. Dokshitzer, Ewarz and J.Nyiri)
7. An Essay On Color Confinement by K.Nishijima and M.Chaichian



# 1. Evolution of the Bogoliubov Renormalization Group <sup>\*</sup>

D.V. Shirkov <sup>†</sup>

N.N.Bogoliubov Laboratory, JINR, Dubna, Russia

## Abstract

We start with a simple introduction into the renormalization group (RG) in quantum field theory and give an overview of the renormalization group method. The third section is devoted to essential topics of the renorm-group use in the QFT. Here, some fresh results are included.

Then we turn to the remarkable proliferation of the RG ideas into various fields of physics. The last section summarizes an impressive recent progress of the “QFT renormalization group” application in mathematical physics.

## Contents

<b>1</b>	<b>Renormalization group primer</b>	<b>26</b>
1.1	Mathematical preliminaries . . . . .	26
1.1.1	<i>Renorm-group folklore</i> . . . . .	26
1.1.2	<i>Group Functional Equation</i> . . . . .	27
1.1.3	<i>Abstract Formulation</i> . . . . .	28
1.2	Definition of the Renorm-Group . . . . .	28
1.2.1	<i>The RG transformation</i> . . . . .	28
1.2.2	<i>Simple generalizations</i> . . . . .	29
1.3	Early history and the RG method . . . . .	30
1.3.1	<i>Renormalization and renormalization invariance</i> . . . . .	30
1.3.2	<i>The discovery of the renormalization group</i> . . . . .	31
1.3.3	<i>Creation of the RG method</i> . . . . .	32
1.4	RG in QED . . . . .	33
1.4.1	<i>Effective Electron Charge</i> . . . . .	33
1.4.2	<i>RG transformations</i> . . . . .	34
<b>2</b>	<b>Renormalization group method</b>	<b>35</b>
2.1	Basic idea . . . . .	35
2.2	Differential formulation . . . . .	36
2.3	General solution . . . . .	37
2.4	RGM algorithm . . . . .	38
2.4.1	<i>Technology of RG Method</i> . . . . .	38
2.4.2	<i>RGM usage in QFT</i> . . . . .	39

<sup>\*</sup>Dedicated to the memory of Nicolaj N. Bogoliubov on the occassion of his 90th birthday.

<sup>†</sup>E.mail: shirkovd@thsun1.jinr.ru

<b>3</b>	<b>RG in QFT</b>	<b>40</b>
3.1	UV analysis in general . . . . .	40
3.1.1	<i>One-coupling case</i> . . . . .	40
3.1.2	<i>Multi-coupling case</i> . . . . .	41
3.2	Perturbative approach to the UV asymptote . . . . .	42
3.2.1	<i>Structure of RG results</i> . . . . .	42
3.2.2	<i>The ghost-pole trouble</i> . . . . .	42
3.2.3	<i>Scalar quartic model</i> . . . . .	43
3.3	Mass-dependent analytic solution . . . . .	45
3.4	Some important results . . . . .	46
<b>4</b>	<b>RG expansion</b>	<b>48</b>
4.1	Critical phenomena . . . . .	48
4.1.1	<i>Spin lattices</i> . . . . .	48
4.1.2	<i>Polymer theory</i> . . . . .	49
4.1.3	<i>Turbulence</i> . . . . .	50
4.2	Paths of RG expansion . . . . .	50
4.3	Two faces of RG in QFT . . . . .	51
<b>5</b>	<b>RG symmetry in mathematical physics</b>	<b>51</b>
5.1	Functional self-similarity . . . . .	51
5.2	Recent application to boundary value problem . . . . .	52
	<b>Bibliography</b>	<b>54</b>

# 1 Renormalization group primer

## 1.1 Mathematical preliminaries

### 1.1.1 Renorm-group folklore

Let us start with some simple statements which can be supposed to be widely known by particle theorists. In the quantum field theory (QFT) the renormalization group (RG) is usually associated with a possibility of presenting any physical quantity,  $F(Q^2, g)$ , calculated under a definite renormalization prescription in the form  $F(Q^2/\mu^2, g_\mu)$  (for simplicity – in a massless case) with the renormalized coupling constant  $g_\mu$  definition attached to some renormalization point (or reference momentum scale)  $Q = \mu$ . Differential RG equation is usually said to be driven from the condition that  $F$  does not depend on the choice of  $\mu$ ,

$$\frac{dF}{d\mu} = 0 \quad . \quad (1)$$

The coupling constant  $g_\mu$  dependence on  $\mu$  is described by a specific function  $\bar{g}(Q^2)$  known as an *effective coupling* (sometimes – effective coupling constant)  $g_\mu = \bar{g}(\mu^2)$ .

Eq.(1) can be written down in the form of a partial linear differential equation (DE)

$$\left[ x \frac{\partial}{\partial x} - \beta(g) \frac{\partial}{\partial g} \right] F(x, g) = 0 \quad (2)$$

where  $x = Q^2/\mu^2$ ,  $g$  stands for  $g_\mu$  and  $\beta(g)$ , the group generator, usually referred to as *beta function*, is defined by

$$\beta(g_\mu) = z \frac{\partial \bar{g}(z)}{\partial z} \quad \text{at} \quad z = \mu^2 \quad .$$

The effective coupling  $\bar{g}$  should be considered as a function of two arguments:  $x = Q^2/\mu^2$  and  $g_\mu$  with the boundary condition  $\bar{g}(1, g) = g$ . Besides

$$\left[ x \frac{\partial}{\partial x} - \beta(g) \frac{\partial}{\partial g} \right] \bar{g}(x, g) = 0 \quad (3)$$

it satisfies the nonlinear DE

$$x \frac{\partial \bar{g}(x, g)}{\partial x} = \beta(\bar{g}(x, g)) \quad (4)$$

which is nothing else but a characteristic equation for (2). To employ this formalism, one has to give  $\beta(g)$ . Usually, for this one uses renormalized perturbation theory.

The foregoing can be considered as a “RG folklore”. For brevity, we gave it in the simplest *massless* version, which corresponds to the UV case, for the QFT model with one coupling constant.

### 1.1.2 Group Functional Equation

Less popular are the RG Functional Equations (FEs). The FE for the  $\bar{g}$  in the UV case has the form

$$\bar{g}(x, g) = \bar{g} \left( \frac{x}{t}, \bar{g}(t, g) \right). \quad (5)$$

This equation, which follows (see, e.g., the Chapter *Renormalization Group* in Ref.[1]) from finite Dyson renormalization transformations, represents a basement of the differential RG formulation. Popular DE (4) can be directly obtained from it by differentiating over  $x$  and then putting  $t = x$ . On the other hand, by differentiating (5) with respect to  $t$  at  $t = 1$  we get partial DE (3).

The FE (5) as well as similar FEs for propagators and vertex functions (see, below, eq.(19)) must be considered as the most compact and general formulation of the RG symmetry in QFT.

However, in reality, group FEs, like (5) (and DEs (4) and (3) as well) do not contain any physics at all being just the reflection of the group composition law! Here, we mean the continuous group (that is, the Lie group of transformations) of operations changing the reference point  $\mu$  involved into the coupling constant  $g_\mu$  definition. Namely, we can regard the change of a reference coupling  $g_\mu \rightarrow g_{\bar{\mu}}$  as an operation of the group element  $T_t$

$$T_t g_\mu = g_{\mu\sqrt{t}} = \bar{g}(t, g_\mu)$$

with a real continuous positive numerical parameter  $t (= \mu^2/\bar{\mu}^2)$ .

If we set  $x = \tau t$ , then the l.h.s. of (5) can be achieved from  $g$  by operation  $T_{\tau t}$ , while the r.h.s. may be identified as  $T_\tau T_t g$ . The content of eq.(5) is just the group composition law,

$$T_{\tau t} = T_\tau T_t.$$

Thus, the essence of the basic RG functional equation (5) is the necessary condition for transformations  $T_t$  to form a group.

At the same time, it demonstrates that function  $\bar{g}$  is *invariant* with respect to simultaneous transformation

$$R_t : \{ x' = x/t, g' = \bar{g}(t, g) \}. \quad (6)$$

Invariance condition for an observable now can be written down as

$$F(x, g) = F \left( \frac{x}{t}, \bar{g}(t, g) \right).$$

Usually, of interest are also functions  $\phi(x, g)$  (like, e.g., propagator amplitudes in QFT) transforming as a linear representation of RG

$$\phi(x, g) \rightarrow R_t \phi = \phi(x', g') = z(t, g) \phi(x, g). \quad (7)$$

Note also that the group FE for an observable, like matrix element, is of the form

$$M(\{x\}, y; g) = M \left( \left\{ \frac{x}{t} \right\}, \frac{y}{t}, g(t, y; g) \right); \quad \{x\} = x_1, x_2, \dots, x_k \quad (8)$$

which reflects an existence of several  $Q^2$ -type arguments and implements its independence of renormalization details corresponding to Eq.(1).

### 1.1.3 Abstract Formulation

To make this point clearer, let us show that FE (5) can formally be obtained directly from the group composition law. Generally, the mathematical formulation of the RG transformation can be presented as a functional realization of the mentioned Lie group.

Consider the transformation  $T(l)$  of a certain abstract set  $\mathcal{M}$  of elements  $M_i$  into itself depending on a continuous real parameter  $l$  ( $-\infty < l < \infty$ ) such that for each element  $M$  we have

$$T(l)M = M' \quad (M, M' \in \mathcal{M}) .$$

Suppose that  $\mathcal{M}$  can be projected onto the real axis, i.e., for every  $M_i$  there corresponds a real number  $g_i$ <sup>1</sup>. Then, this transformation can be written in the analytic form

$$T(l)g = g' = G(l, g) ,$$

$G$  being a continuous function of two arguments satisfying the normalization condition  $G(0, g) = g$  which corresponds to the identity transformation  $T(0) = \mathbf{E}$ .

Transformations  $T(l)$  form a group if they satisfy the composition law

$$T(l) \times T(\lambda) = T(l + \lambda) \quad (9)$$

to which there corresponds the functional equation for  $G$  :

$$G\{l, G(\lambda, g)\} = G(l + \lambda, g) . \quad (10)$$

As it follows from the bases of the Lie group theory, it is sufficient to deal with the infinitesimal transformation at  $\lambda \ll 1$ , i.e., with the DE

$$\frac{\partial G(l, g)}{\partial l} = \beta\{G(l, g)\} . \quad (11)$$

Here the group generator is defined as

$$\beta(g) = \frac{\partial G(\epsilon, g)}{\partial \epsilon} \quad \text{at} \quad \epsilon = 0 .$$

Performing a logarithmic change of variables

$$l = \ln x, \quad \lambda = \ln t, \quad G(l, g) = \bar{g}(x, g), \quad T(\ln t) = T_t \quad (12)$$

we obtain (5) and (4) instead of (10) and (11).

## 1.2 Definition of the Renorm-Group

### 1.2.1 The RG transformation

Generally, the RG can be defined as a continuous one-parameter group of specific transformations of a partial solution (or the solution characteristic) of a problem, a solution that is fixed by a boundary condition. The RG transformation involves boundary condition parameters and corresponds to some change in the way of imposing this condition.

For illustration, imagine an one-argument solution characteristic  $f(x)$  that has to be specified by the boundary condition  $f(x_0) = f_0$ . Formally, represent the given characteristic of a partial solution as a function of boundary parameters as well:  $f(x) = f(x, x_0, f_0)$ . (This step can be considered as an *embedding* operation). The RG transformation then corresponds to a changeover of the way of parameterization, say from  $\{x_0, f_0\}$  to  $\{x_1, f_1\}$  for the *same* solution. In other words, the  $x$  argument value, at which the boundary condition is given, does not need to be  $x_0$ , but we may choose another point  $x_i$ . Our solution  $f$  can be written in a form of a two-argument function

<sup>1</sup>This condition is not essential and can be modified — see, below, eqs. (16) and (17).

$F(x/x_0, f_0)$  with the property  $F(1, \gamma) = \gamma$ . The equality  $F(x/x_0, f_0) = F(x/x_1, f_1)$  reflects the fact that under such a change of a boundary condition the form of function  $F$  itself is not modified (as, e.g., in the case of  $F(x, \gamma) = \Phi(\ln x + \gamma)$ ). Noting that  $f_1 = F(x_1/x_0, f_0)$ , we obtain

$$F(\xi, f_0) = F(\xi/t, F(t, f_0)) \quad ; \quad \xi = x/x_0, \quad t = x_1/x_0.$$

The group transformation here is  $\{ \xi \rightarrow \xi/t, \quad f_0 \rightarrow f_1 = F(t, f_0) \}$ .

The renorm-group transformation for a given solution of some physical problem in the simplest case can be defined as

a simultaneous one-parameter transformation of two variables, say  $x$  and  $g$ , by

$$R_t : \{ x \rightarrow x' = x/t, \quad g \rightarrow g' = \bar{g}(x, g) \}, \quad (6)$$

the first being a scaling of a coordinate  $x$  and the second — a more complicated functional transformation of the solution characteristics. Eq.(5) for the transformation function  $\bar{g}$  provides the group property  $T_{\tau t} = T_{\tau} T_t$  of the transformation (6). Performing the logarithmic change of variables and an appropriate redefinition of a transformation function (12), we obtain eqs.(10), (11) and

$$R(l) : \{ q \rightarrow q' = q - l, \quad g \rightarrow g' = G(l, g) \}, \quad (13)$$

instead of (5), (4) and (6). One can refer to these equations as the *multiplicative* version (and previous equations in abstract formulation as to the *additive* one). They are just the RG equations and transformation for a massless QFT model with one coupling constant. In that case  $x = Q^2/\mu^2$  is the ratio of a 4-momentum  $Q$  squared to a “normalization” momentum  $\mu$  squared and  $g$ , the coupling constant.

Several generalizations of (6) and (13) will be considered below.

### 1.2.2 Simple generalizations

#### “Massive” Case

For example in QFT, if we do not neglect the mass  $m$  of a particle, we have to insert an additional dimensionless argument into the invariant coupling  $\bar{g}$  which now has to be considered as a function of three variables:  $x = Q^2/\mu^2$ ,  $y = m^2/\mu^2$ , and  $g$ . The presence of a new “mass” argument  $y$  modifies the group transformation

$$R_t : \left\{ x' = \frac{x}{t}, \quad y' = \frac{y}{t}, \quad g' = \bar{g}(t, y; g) \right\} \quad (14)$$

and the functional equation

$$\bar{g}(x, y; g) = \bar{g} \left( \frac{x}{t}, \frac{y}{t}; \bar{g}(t, y; g) \right). \quad (15)$$

Here, it is important that the new parameter  $y$  (which in physical nature must be close to the  $x$  variable as it scales similarly) enters also into the transformation law of  $g$ .

If the considered QFT model, like QCD, contains several masses there will be several mass arguments

$$y \rightarrow \{y\} = y_1, y_2, \dots, y_n.$$

#### Multi-coupling case

A more complicated generalization corresponds to transition to the case with several coupling constants:  $g \rightarrow \{g\} = g_1, \dots, g_k$ . Here, one has to introduce the “family” of effective couplings

$$\bar{g} \rightarrow \{\bar{g}\}, \quad \bar{g}_i = \bar{g}_i(x, y; \{g\}); \quad i = 1, 2, \dots, k,$$



satisfying the system of coupled functional equations

$$\bar{g}_i(x, y; \{g\}) = \bar{g}_i\left(\frac{x}{t}, \frac{y}{t}; \{\bar{g}(t, y; \{g\})\}\right). \quad (16)$$

In the abstract formulation this system is a generalization of (5) and (15) for the case when every element  $M_i$  of  $\mathcal{M}$  can be described by  $k$  numerical parameters, i.e., by a point  $\{g\}$  in the  $k$ -dimensional real parameter space. The RG transformation now is

$$R_t : \left\{ x \rightarrow \frac{x}{t}, \quad y \rightarrow \frac{y}{t}, \quad \{g\} \rightarrow \{g(t)\} \right\}, \quad g_i(t) = \bar{g}_i(t, y; \{g\}). \quad (17)$$

### 1.3 Early history and the RG method

#### 1.3.1 Renormalization and renormalization invariance

As it is known, the regular formalism for eliminating the UV divergences in QFT was developed on the basis of covariant perturbation theory for the scattering  $S$  matrix in the late 40s. This breakthrough is connected with the names of Tomonaga, Feynman, Schwinger and some others. In particular, Dyson and Abdus Salam carried out the general analysis of the structure of divergences in arbitrarily high orders of perturbation theory. Nevertheless, a number of subtle questions concerning overlapping divergences remained unclear.

An important contribution in this direction based on a thorough analysis of the mathematical nature of UV divergences was made by Bogoliubov. This was achieved on the basis of a branch of mathematics which was new at that time, namely, the Sobolev Schwartz *theory of distributions*. The point is, that propagators in local QFT are distributions (similar to the Dirac delta function) and propagator products appearing in the coefficients of the  $S$  matrix expansion require a supplementary definition in the case when their arguments coincide and lie on the light cone.

In the mid 50s on the basis of this approach Bogoliubov and his disciples developed a technique of supplementing the definition of products of singular Stückelberg–Feynman propagators [2] and proved a theorem [3] on the finiteness and uniqueness (for renormalizable theories) of the  $S$  matrix elements in any order of perturbation theory. The prescriptive part of this theorem, the *Bogoliubov  $R$ -operation* (see, e.g., chapter “Removal of divergencies from the  $S$  matrix” in the monograph [1]), still remains a practical means of obtaining finite and unique results in the higher order perturbation calculation.

The Bogoliubov algorithm works, essentially, as follows:

- To remove the UV divergences of a one-loop diagram, instead of introducing some regularization, e.g., the momentum cutoff, and handling (quasi) infinite counterterms, it suffices to complete the definition of a divergent Feynman integral by subtracting from it a certain polynomial in the external momenta which in the simplest case is reduced to the first few terms of the Taylor series.
- For multi-loop diagram (including one with overlapping divergences) one should first subtract all divergent subdiagrams and finish with subtracting the diagram as a whole in a hierarchical order regulated by the  $R$ -operator.

An attractive feature of this approach is that it is free from any auxiliary nonphysical attributes such as bare masses, bare coupling constants, and regularization parameters which are not involved in the computation within the Bogoliubov’s algorithm. The latter can be regarded as *renormalization without regularization and counterterms*.

The uniqueness of computational results for the observable  $S$ -matrix elements is ensured by special conditions imposed on them. These conditions contain some degree of freedom (related to different renormalization schemes and momentum scales) that can be used to establish finite relations between the Lagrangian parameters (masses, coupling constants) and corresponding renormalized quantities. The fact that physical predictions are independent of arbitrariness in the renormalization conditions, that is, they are *renorm-invariant*, constitutes the conceptual foundation of the renormalization group.

### 1.3.2 The discovery of the renormalization group

In the 1952-1953 Stückelberg and Peterman [5] discovered<sup>2</sup> a group of infinitesimal transformations related to finite arbitrariness arising in the  $S$ -matrix elements upon elimination of the UV divergences. These authors introduced *normalization group* generated by Lie operators connected with renormalization of the coupling constant  $e$ .

In the following year, on the basis of (infinite) Dyson's renormalization transformations formulated in the regularized form, Gell-Mann and Low [6] derived functional equations for the QED propagators in the UV limit. The appendix to this article contains the general solution (obtained by T.D. Lee) of this functional equation for the renormalized transverse photon propagator amplitude  $d(x, e^2)$ , written in two equivalent forms:

$$e^2 d(x, e^2) = F(x F^{-1}(e^2)) \quad \text{and} \quad \ln x = \int_{e^2}^{e^2 d} \frac{dy}{\psi(y)}, \quad \psi(e^2) = \left. \frac{\partial(e^2 d)}{\partial \ln x} \right|_{x=1}. \quad (18)$$

A qualitative analysis of the behaviour of the quantum electromagnetic interaction at small distances was carried out with the aid of (18). Two possibilities, namely, infinite and finite charge renormalizations were pointed out.

However, paper [6] paid no attention to the group character of the analysis and the results obtained there. The authors missed a chance to establish a connection between their results and the standard perturbation theory and did not discuss the possibility that a ghost pole solution might exist.

The final step was taken by Bogoliubov and the present author bs-55a,bs-55b,sh-55 — see also the survey [10] published in English in 1956. Using the group properties of finite Dyson transformations for the coupling constant, fields and Green functions, these authors derived functional group equations for the propagators and vertices in QED in the general case (that is, with the electron mass taken into account). For example, the equation for the transverse amplitude of the photon propagator and electron propagator amplitude were obtained in the form

$$\begin{aligned} d(x, y; e^2) &= d(t, y; e^2) d\left(\frac{x}{t}, \frac{y}{t}; e^2 d(t, y; e^2)\right), \\ s(x, y; e^2) &= s(t, y; e^2) s\left(\frac{x}{t}, \frac{y}{t}; e^2 d(t, y; e^2)\right) \end{aligned} \quad (19)$$

in which the dependence on the mass variable  $y = m^2/\mu^2$  was present.

As can be seen, the product  $e^2 d$  of electron charge squared and photon propagator amplitude enters in both the FEs. This product is invariant with respect to Dyson's transformation. We called this function — *invariant charge* and introduced the term *renormalization group*.

In the modern notation, the first equation is that for the invariant charge (now widely known as an effective or running coupling)  $\bar{\alpha} = \alpha d(x, y; \alpha = e^2)$ :

$$\bar{\alpha}(x, y; \alpha) = \bar{\alpha}\left(\frac{x}{t}, \frac{y}{t}; \bar{\alpha}(t, y; \alpha)\right). \quad (20)$$

Let us emphasize that, unlike the approach Ref.[6], in the latter case there is no relation with UV divergences and simplification due to the massless nature of the UV asymptotics. Here, the homogeneity of the transfer momentum scale is violated explicitly by mass  $m$ . Nevertheless, the symmetry (even though a bit more complex one) underlying the RG, as before, can be stated as an *exact symmetry* of the solutions of the QFT problem. This is what we mean when using the term Bogoliubov's renormalization group or *renorm-group* for short.

The differential group equations for  $\bar{\alpha}$  and for the electron propagator:

$$\frac{\partial \bar{\alpha}(x, y; \alpha)}{\partial \ln x} = \beta\left(\frac{y}{x}, \bar{\alpha}(x, y; \alpha)\right); \quad \frac{\partial s(x, y; \alpha)}{\partial \ln x} = \gamma\left(\frac{y}{x}, \bar{\alpha}(x, y; \alpha)\right) s(x, y; \alpha), \quad (21)$$

<sup>2</sup>For a more detailed exposition of the RG early history see our review [4].

with

$$\beta(y, \alpha) = \frac{\partial \bar{\alpha}(\xi, y; \alpha)}{\partial \xi}, \quad \gamma(y, \alpha) = \frac{\partial s(\xi, y; \alpha)}{\partial \xi} \quad \text{at } \xi = 1 \quad (22)$$

were first derived in [7] by differentiating the FEs. In this way, explicit realization of the group DEs mentioned in the paper [5] was obtained. These results established a conceptual link with the Stückelberg—Peterman and Gell-Mann—Low results.

### 1.3.3 Creation of the RG method

Another important achievement of paper [7] consisted in formulating a simple algorithm for improving an approximate perturbative solution by combining it with the Lie equations (for detail, see below, Section 2).

In our adjacent publication [8] this algorithm was effectively used to analyse the UV and infrared (IR) behaviour in QED. The one-loop and two-loop UV asymptotics

$$\bar{\alpha}_{RG}^{(1)}(x; \alpha) \equiv \bar{\alpha}_{RG}^{(1)}(x, 0, \alpha) = \frac{\alpha}{1 - \frac{\alpha}{3\pi} \ln x}, \quad (23)$$

$$\bar{\alpha}_{RG}^{(2)}(x; \alpha) = \frac{\alpha}{1 - \frac{\alpha}{3\pi} \ln x + \frac{3\alpha}{4\pi} \ln(1 - \frac{\alpha}{3\pi} \ln x)} \quad (24)$$

of the photon propagator as well as the IR behavior

$$s(x, y; \alpha) \approx (x/y - 1)^{-3\alpha/2\pi} \approx (p^2/m^2 - 1)^{-3\alpha/2\pi} \quad (25)$$

of the electron propagator in the transverse gauge were obtained. At that time, these expressions were already known only at the one-loop level. It should be noted that in the mid 50s the problem of the UV behaviour in local QFT was quite urgent. At that time, substantial progress in the analysis of QED at small distances was made by Landau and his collaborators [11]. However, Landau's approach did not provide a prescription for constructing subsequent approximations.

The simple technique for obtaining higher approximations was found only within the new renorm-group method. The one-loop UV asymptotics of QED propagators obtained in our paper [8], eqs. (23) and (25), agreed precisely with the results of Landau's group.

Within the RG approach these results can be obtained in just a few lines of argumentation. To this end, e.g., the massless one-loop perturbation approximation should be substituted into the r.h.s. of the first equation in (22) to compute the generator  $\beta(0, \alpha) = \psi(\alpha) = \alpha^2/3\pi$  followed by elementary integration of the first of eqs.(21).

Moreover, starting from the next order perturbation expression  $\bar{\alpha}_{PTH}^{(2)}(x; \alpha)$  containing the  $\alpha^3 \ln x$  term, we arrived at the second renorm-group approximation (24) performing summation of the  $\alpha^2(\alpha \ln)^n$  terms. This two-loop solution for the invariant coupling first obtained in [8] contains the nontrivial log-of-log dependence which is now widely known of the “next-to-leading logs” approximation for the running coupling in quantum chromodynamics (QCD) — see, below, eq.(47).

Comparing solution (24) with (23), one can conclude that the two-loop correction is extremely essential just in the vicinity of the ghost pole singularity at  $x_1 = \exp(3\pi/\alpha)$ . This demonstrates that the RG method is a regular procedure, within which it is quite easy to estimate the range of applicability of the results.

Quite soon, this approach was formulated [9] for the case of QFT with two coupling constants, say,  $g$  and  $h$ , namely, for a model of pion-nucleon interactions with the pions self-interaction. To the system of functional equations for two invariant couplings

$$\begin{aligned} \bar{g}^2(x, y; g^2, h) &= \bar{g}^2\left(\frac{x}{t}, \frac{y}{t}, \bar{g}^2(t, y; g^2, h), \bar{h}(t, y; g^2, h)\right), \\ \bar{h}(x, y; g^2, h) &= \bar{h}\left(\frac{x}{t}, \frac{y}{t}, \bar{g}^2(t, y; g^2, h), \bar{h}(t, y; g^2, h)\right) \end{aligned}$$

there corresponds a coupled system of nonlinear DEs – see, below, eqs.(52). It was analysed [12] in the one-loop approximation to carry out the UV analysis of the renormalizable model of the pion-nucleon interaction.

In a more general case of arbitrary covariant gauge the RG analysis in QED was carried out in [13]. Here, the point was that the charge renormalization is connected only with the transverse part of the photon propagator. Therefore, under nontransverse covariant (e.g., Feynman) gauge the Dyson transformation has a more complex form. This issue has been resolved by considering the gauge parameter as another coupling constant.

In Refs.[7, 8, 9, 13] and [12] the RG was thus directly connected with practical computations of the UV and IR asymptotics. Since then this technique, known as the *renormalization group method* (RGM) and being summarized in the first edition of monograph [1], has become the sole means of asymptotic analysis in local QFT.

## 1.4 RG in QED

### 1.4.1 Effective Electron Charge

An essential feature of quantum theory is the presence of virtual states and transitions. In QED, e.g., the process of virtual dissociation of a photon into an electron-positron pair and vice versa  $\gamma \leftrightarrow e^+ + e^-$  can take place. The sequence of two such virtual transitions represents the simplest contribution to the effect of vacuum polarization.

The vacuum polarization processes lead to several specific phenomena and particularly to the notion of *effective electron charge*. To explain this, let us start with a classical analogy.

Take a polarizable medium consisting of molecules that can be imagined as electric dipoles. Insert into it an external electric charge  $Q$ . Due to the attraction of opposite charges, the dipoles change their position so that the charge  $Q$  turns out to be partially screened. As a result, at a distance  $r$  from  $Q$  the electric potential will be smaller than the vacuum Coulomb law  $Q/r$  and can be presented in the form  $Q(r)/r$  where, generally,  $Q(r) \leq Q$ . The introduced quantity  $Q(r)$  is known as an effective charge. As  $r$  decreases,  $Q(r)$  increases and as  $r \rightarrow 0$ ,  $Q(r)$  tends to  $Q$ .

In QFT the vacuum, i.e., the interparticle space itself stands for the “polarizable medium”. Quantum-field vacuum is not physically empty. It is filled with vacuum fluctuations, i.e., with virtual particles. These “zero fluctuations” are a well-known effect of a ground state in quantum world. In QED, zero oscillations consist mainly of short-lived virtual ( $e^+, e^-$ ) pairs which play the role of tiny electric dipoles.

Consider the process of measuring the electron charge with the help of some external electromagnetic field. In the quantum case the probing photon can virtually dissociate into the ( $e^+, e^-$ ) pair. This pair can be treated as a virtual dipole that produces partial screening of the measured charge. The simplest process involves two elementary electromagnetic interactions, its contribution to an effective charge being proportional to the small number  $e^2 \equiv \alpha \simeq 1/137$ ; and this contribution depends on the distance  $r$ ! In the region of  $r$  values much smaller than the Compton length of the electron  $r_e = h/mc \simeq 3,9 \cdot 10^{-11} \text{cm}$  it depends on  $r$  logarithmically

$$e \rightarrow e(r) = e \left\{ 1 - \frac{\alpha}{3\pi} \ln \frac{r}{r_e} + \dots \right\} \quad (26)$$

as was first discussed by Dirac [14] in the middle of the 30s. The  $e(r)$  value decreases as  $r$  grows. So, qualitatively, the QED effective charge behavior corresponds to a classical picture of screening.

This dependence can be presented by a set of curves  $e(r)$ . Each curve represents a possible behavior of the effective charge  $e(r)$  as obtained from the theory and considered without any reference to experiment ( $\alpha = e^2$  being unspecified numerically).

The point is that in the classical analogue the value of an external charge  $Q$  inserted into the polarizable medium is known from the very beginning. In quantum physics it is not the case, and a charge value can be measured at *not very small* distances. The result of measurement generally has to be specified by two quantities: the “distance of measurement”  $r_i$  and the measured charge value  $e_i$ . Hence, to make the choice from the mentioned set of curves, one has to fix the point on

the plane with the coordinates  $r = r_i$ ,  $e(r) = e_i$ . Thus, for the chosen “physical” curve  $e(r_i) = e_i$ . Note, that the usual definition of the electron charge by a classical macroscopic (like Millikan) experiment corresponds here to very large distances  $r \geq r_e$ , i.e.,  $e = e(r = r_e) = 1/\sqrt{137}$ .

As it is well known, in relativistic microphysics one usually uses the momentum rather than coordinate representation. Correspondingly, instead of  $e(r)$  one deals with the quantity  $\alpha(Q^2)$ , the Fourier transform of  $e(r)$  squared. It is a monotonically increasing function of its argument  $Q^2$ , the 4-momentum transfer squared. Here and below, the bar denotes a function (distinct from  $\alpha$ ,  $\alpha_\mu$ ,  $\alpha_i$  – its numerical values at some given value of the  $Q^2$  argument). The correspondence condition with the classical electrodynamics now takes the form  $\alpha(0) = 1/137$ , as in our scale the external long-range field corresponds to a photon with vanishing 4-momentum. However, as before, to fix one of possible curves on the plane  $(Q, \bar{\alpha})$  one has to give a point  $Q \equiv \sqrt{Q^2} = \mu$ ,  $\alpha \equiv \alpha_\mu$  and hence, for the selected curve  $\bar{\alpha}(\mu^2) = \alpha_\mu$ .

The parameter  $\mu$  sometimes is referred to as a *scale parameter*. As is clear, it is just the momentum magnitude for a photon used for the charge measurement. The effective coupling function  $\bar{\alpha}(Q^2)$  describes the dependence of the electron charge value on the measurement conditions. In our days the logarithmic corrections to the Millikan value become essential and are measured at big accelerators.

The parameter  $\mu$  has no analogue in the QED Lagrangian that reproduces the classical electrodynamical one. The phenomenon of its arising in QFT was exaggerated by the term “dimensional transmutation”. As it was shown, its appearance is very natural and is connected with the measurement procedure.

This is a good place to recall the ideas by Niels Bohr formulated in the middle of 30s [15] and related to the complementarity principle. The point is that to specify a quantum system, it is necessary to fix its “macroscopic surrounding”, i.e., to give the properties of macroscopic devices used in the measurement process. Just these devices are described by additional parameters, like  $\mu$ . However, this is not the end of the Bohr (i.e. – scale) parameter story. As can be shown, in the QFT this parameter existence leads to a new symmetry lying in the foundation of the renormalization group.

#### 1.4.2 RG transformations

To do this, consider again the function  $\bar{\alpha}(Q^2)$  having in mind that the physical solution has been chosen by the condition  $\bar{\alpha}(Q^2 = \mu^2) = \alpha_\mu$ . Assume also for simplicity that we deal with a massless QED, more precisely, with the approximation  $|Q| \gg m$ . This corresponds to the GeV-energy region or to distances  $r \ll r_e$ . Here the effective charge function  $\bar{\alpha}$  can be represented as a function of two dimensionless arguments  $Q^2/\mu^2$  and  $\alpha_\mu$ , i.e.,  $\bar{\alpha}(Q^2) = \bar{\alpha}(Q^2/\mu^2, \alpha_\mu)$ .

Now, take into account that the couple of parameters  $\mu, \alpha_\mu$  used to identify the physical solution may, generally, correspond to any scale  $\mu$ . Take two scales “1” and “2” with coordinates  $\mu_1, \alpha_1$  and  $\mu_2, \alpha_2$ , respectively. It is evident that  $\bar{\alpha}$  can be parameterized by any pair  $\mu_i, \alpha_i$ ;  $i = 1, 2, \dots$  so that for arbitrary  $Q^2$  values the identity

$$\bar{\alpha}(Q^2/\mu_1^2, \alpha_1) = \bar{\alpha}(Q^2/\mu_2^2, \alpha_2)$$

should hold.

At the same time the second argument in the r.h.s.,  $\alpha_2$ , which by definition is equal to  $\bar{\alpha}$  at  $Q^2 = \mu_2^2$ , can be expressed in terms of  $\bar{\alpha}$  parameterized with the help of point “1” coordinates, i.e.,  $\alpha_2 = \bar{\alpha}(Q^2 = \mu_2^2) = \bar{\alpha}(\mu_2^2/\mu_1^2, \alpha_1)$ . Combining the last two relations and introducing a notation  $Q^2/\mu^2 = x$ ,  $\alpha_1 = \alpha$ ,  $\mu_2^2/\mu_1^2 = t$ , we arrive at the FE

$$\bar{\alpha}(x, \alpha) = \bar{\alpha}\left(\frac{x}{t}, \bar{\alpha}(t, \alpha)\right), \quad (27)$$

identical with eq.(5).

Note, that the corresponding continuous one-parameter transformation is just the change (“1”  $\rightarrow$  “2”) of the parameterization point

$$R_t : \{\mu_1 \rightarrow \mu_2 = \sqrt{t}\mu_1, \alpha_1 \rightarrow \alpha_2 = \bar{\alpha}(t, \alpha_1)\} \quad (28)$$

Thus, we have shown that in the renormalized QED there exists invariance with respect to continuous transformations of the group type which involve two quantities and contain a functional dependence. This precisely corresponds to definition (6).

As can be shown, in QED the effective coupling  $\bar{\alpha}$  is equal to a product of  $\alpha$  and dimensionless function  $d(x, \alpha)$  – the transverse photon propagator amplitude with due regard for vacuum polarization effects. Generally, in QFT models with one coupling constant the invariant coupling  $\bar{g}(x, g)$  can be expressed as a product of  $g$ , corresponding vertex function and the square root of propagator amplitudes of the fields participating in the interaction. Usually, this can be done on the basis of Dyson finite renormalization transformations.

Thus, the RG invariance is nothing else but the invariance of a solution with respect to the way of its parameterization. For instance, in real QED, instead of using the “Millikan’s value”  $\bar{\alpha}(0) = 1/137$  one may take the “CERN value”  $\bar{\alpha}(M_Z^2) \simeq 1/128,9$ .

## 2 Renormalization group method

### 2.1 Basic idea

Approximate solution of the physical problems with RG symmetry usually does not obey this symmetry which is lost in the course of approximation. This is essential when the solution under consideration possesses a singularity as far as the singularity structure commonly is destroyed by an approximation.

In QFT, e.g., the usual way of calculation is based on the perturbation method, i.e., on power expansion in  $g$ . It is not difficult to see that finite sums of this expansion do not satisfy the functional group equations. As the simplest illustration, consider the effective coupling  $\bar{g}$  in the UV region where the one-loop contribution has a logarithmic form

$$\bar{g}_{\text{PT}}^{(1)}(x, g) = g + g^2 \beta \ln x \quad (29)$$

with  $\beta$ , a numerical coefficient. By substituting this expression into FE (5), after simple manipulation one has

$$\begin{aligned} \text{Discr}[\bar{g}_{\text{PT}}^{(1)}] &\equiv \bar{g}_{\text{PT}}^{(1)}(x, g) - \bar{g}_{\text{PT}}^{(1)}\left(x/t, \bar{g}_{\text{PT}}^{(1)}(t, g)\right) = \\ &= [g + g^2 \beta \ln x] - [g + g^2 \beta \ln x + 2g^3 \beta^2 \ln t \ln(x/t)] \neq 0 \end{aligned}$$

— error in the  $g^3$  order. This discrepancy can be liquidated by addition of the particular next order term to the r.h.s. of (29)

$$\bar{g}_{\text{PT}}^{(2)} = g + g^2 \beta \ln x + g^3 \beta^2 \ln^2 x.$$

This “improved” expression would yield the discrepancy of the  $g^4$  order which in its turn can be abolished by adding the  $g^4 \ln^3$  term to (29) and so on.

Thus, we see, on the one hand, that the finite polynomials cannot satisfy the condition of renormalization invariance. On the other hand, we can conclude that the functional RG equation represents a tool for iterative reconstruction of renorm-invariant expression that has the form of infinite series.

This example illustrates a rather general situation. As a rule, approximate solutions do not satisfy a group symmetry. Here, this happens in the UV limit as  $\ln x \rightarrow \infty$  where the observed discrepancy becomes important.

Another illustration is provided by the one-dimensional transfer problem (for detail, see our reviews in Refs. [16]). Take a half-space ( $l > 0$ ) filled with a homogenous media. Let some given amount of particles (or radiation) be falling on the surface (at  $l = 0$ ) from the empty half-space. The particle density  $n(l, \mathbf{v})$  is a function of coordinate and of particle velocity  $\mathbf{v}$ . It satisfies an integro-differential kinetic Boltzmann equation. In some cases one can neglect by the energy dependence of cross-sections. Here, the solution can be treated as a function of coordinate and

direction of particle velocity  $\Omega = \mathbf{v}/v$ . In this, “one-velocity”, case the simple symmetry of the RG type was found not for density, but for  $n$  integrated over directions in forward hemisphere

$$G(l) = \int_{\Omega_+} n(l, \Omega) d\Omega .$$

The function  $G$  relates to amount of all particles moving inwards the media. A partial solution of this problem will depend on the boundary condition at  $l = 0$ . Correspondingly, the solution characteristic  $G$  will be the function of two arguments  $G(l, g)$  — coordinate  $l$  (distance from the boundary) and total amount of ingoing particles  $g = G(l = 0) = G(0, g)$ . Just this function  $G(l, g)$  satisfies [17] the group FE (10) in the additive form.

It is rather simple to get an approximate behavior of this density  $G(l, g)$ , at small  $l$

$$G(l, g) = g + lG'(0, g), \quad l \ll 1, \quad (30)$$

which, being considered for large  $l$  values, also does not obey the mentioned symmetry.

On this basis one can set the task of “renormalization-invariant improvement” of perturbative results. The key idea is to combine an approximate solution with the group equations. The simplest and most convenient way for this “marriage” is the use of Lie equations, i.e., group differential equations. The renormalization group method (RGM) as it was first formulated in Refs.[7, 8, 10] is essentially based on these group equations.

## 2.2 Differential formulation

The differential equations can be obtained from the functional ones in two different ways. Differentiating eq.(15) by  $x$  and putting then  $t = x$ , one obtains (compare with (4)):

$$x \frac{\partial \bar{g}(x, y, g)}{\partial x} = \beta \left( \frac{y}{x}, \bar{g}(x, y; g) \right), \quad \beta(y, g) = \left. \frac{\partial g(t, y; g)}{\partial t} \right|_{t=1}. \quad (22a)$$

The nonlinear equation (22a) can be considered as “massive” generalization of eq.(4). On the other hand, one can differentiate eq.(15) with respect to  $t$  at the point  $t = 1$ , which yields

$$\left[ x \frac{\partial}{\partial x} + y \frac{\partial}{\partial y} - \beta(y, g) \frac{\partial}{\partial g} \right] g(x, y; g) = 0, \quad (31)$$

a linear partial differential equation (PDE).

Analogous operations applied to the second of eqs.(19) lead to:

$$\frac{\partial s(x, y; g)}{\partial \ln x} = \gamma \left[ \frac{y}{x}, g(x, y; g) \right] s(x, y; g) \quad (22b)$$

and

$$\left\{ x \frac{\partial}{\partial x} + y \frac{\partial}{\partial y} - \beta(y, g) \frac{\partial}{\partial g} - \gamma(y, g) \right\} s(x, y; g) = 0 \quad (32)$$

where

$$\gamma(y, g) = \left. \frac{\partial s(t, y; g)}{\partial t} \right|_{t=1} \quad (33)$$

is the so-called *anomalous dimension* of  $s$ . For a group invariant, like, e.g., matrix element  $M$  satisfying FE (8) this dimension is equal to zero. The corresponding PDE looks like

$$\left\{ \sum_i x_i \frac{\partial}{\partial x_i} - y \frac{\partial}{\partial y} - \beta(y, g) \frac{\partial}{\partial g} \right\} M(x, y; g) = 0. \quad (34)$$

Equations (31), (32), and (34) express the independence on the  $t$  parameter of the r.h.s. of the related functional group equations, i.e., a mutual compensation of  $t$  dependences via three

(or more) arguments. This DEs can be called *compensational* equations to distinguish them from nonlinear eqs.(21) which can be referred to as *evolutional* group equations.

Stress that compensational as well as evolutional DEs taken together with normalization (i.e. boundary) conditions like  $\bar{g}(1, g) = g$ ,  $s(1, g) = 1$  are equivalent to functional equations and to each other. At the same time, evolutional Lie equations turn out to be more convenient for practical construction of the solution, generators  $\beta, \gamma$  being given.

Let us comment also that the UV limit of compensational DEs like, e.g.,

$$\left\{ x \frac{\partial}{\partial x} - \beta(g) \frac{\partial}{\partial g} - \gamma(g) \right\} s(x, g) = 0 \quad (35)$$

coincides with the UV limit of specific nonclosed equation

$$\left\{ x \frac{\partial}{\partial x} - \beta(g) \frac{\partial}{\partial g} - \gamma(g) \right\} s(x, g) = \Delta S \quad (C - S)$$

obtained in the early 70s by Callan and Symansik. The r.h.s. of this equation contains the result of mass counter-term insertion into all internal lines of all diagrams for the function  $s$  under consideration. For this reason, in current literature compensational equations are often related to as the Callan–Symansik equations. However, these equations just in the form (31) and (32) were first obtained by Lev Ovsyannikov in 1956 while solving [18] functional RG equations. Therefore, we consider it justifiable to relate compensational DEs to the Ovsyannikov's rather than to some other names.

It is not difficult to formulate group DEs for a multi-coupling case by proper differentiation of FEs (16). For instance, the system of evolutional DEs looks like

$$x \frac{\partial \bar{g}_i(x, y, \{g\})}{\partial x} = \beta_i \left( \frac{y}{x}, \{ \bar{g}_j(x, y; \{g\}) \} \right) . \quad (36)$$

## 2.3 General solution

General solution of the group FEs was obtained in the paper [18] by applying the theory of PDE to the compensational eqs. (31) and (32). Details of the derivation can be found in the Section 48.3 of the third edition of the monograph [1]. The results obtained can be formulated as follows:

To every solution of DE (31), there corresponds some function of two arguments  $F(y, g)$ , reversible with respect to its second argument and connected to  $\bar{g}$  by the relation

$$F(y, g) = F \left( \frac{y}{x}, \bar{g}(x, y; g) \right) . \quad (37)$$

The explicit form of  $\bar{g}$  can be obtained now by reversing the r.h.s:

$$\bar{g}(x, y, g) = F_{(2)}^{-1} \left[ \frac{y}{x}, F(y, g) \right] .$$

To determine  $F$  it is sufficient to specify the generator  $\beta(y, g)$ .

Note also that to get from the Ovsyannikov result (37) the solution in the UV limit, i.e., in a massless case at  $y = 0$ , one has to assume for  $F$  a specific limiting form

$$F(y, g) = y \exp[f(g)] \quad \text{or} \quad = \ln y + f(g) \quad \text{as} \quad y \rightarrow 0 .$$

Then

$$f\{\bar{g}(x, g)\} - f(g) = \ln x ; \quad \bar{g} = f^{-1}\{\ln x + f(g)\} . \quad (38)$$

Here,  $f'(g) = 1/\beta(g)$ . This is equivalent to the Gell-Mann—Low—Lee solution (18).

To every solution of eq.(19) for a function  $s$ , there corresponds some function  $\Sigma(y, g)$  related to  $s$  by

$$s(x, y; g) = \frac{\Sigma[y/x, \bar{g}(x, y; g)]}{\Sigma(y, g)} . \quad (39)$$



Let us give also the general solution of the same type for the system (16) for the  $k$ -couplings case. It can be written down in terms of  $k$  arbitrary functions  $F_i$ , reversible simultaneously with respect to last arguments, and defined from the system of  $k$  functional relations

$$F_i(y, \{g\}) = F_i \left[ \frac{y}{x}, \{ \bar{g}(x, y; \{g\}) \} \right], \quad \{g\} = g_1, \dots, g_k; \quad i, j = 1, \dots, k. \quad (40)$$

All solutions (37) -- (40) satisfy the usual normalization conditions.

The transition to the massless limit in expressions (39) -- (40) can be performed by a trick analogous to the given above. Then, e.g.,

$$s(x, g) = x^{\frac{\Sigma\{g(x, g)\}}{\Sigma(g)}}.$$

Let us also formulate solution for the 2-coupling case  $g_1 = g$ ,  $g_2 = h$  in the massless limit in the form analogous to (38)

$$f_i(\bar{g}, \bar{h}) = f_i(g, h) + \ln x, \quad i = 1, 2. \quad (41)$$

From the solutions presented it follows that imposing group symmetry one reduces by unity the number of independent arguments.

## 2.4 RGM algorithm

### 2.4.1 Technology of RG Method

The idea of the approximate solution marriage [7, 8] with group symmetry can be realised with help of group DEs. If we define group generators  $\beta, \gamma$  from some approximate solutions and then solve evolutional DEs, we obtain *RG improved* solutions that obey the group symmetry and correspond to the approximate solutions used as an input.

Now we can formulate an algorithm of improving an approximate solution. The procedure is given by the following recipe which we illustrate by a massless one coupling case (4) and (5):

Assume some approximate solution  $\bar{g}_{\text{appr}}$  is known.

1. On the basis of eq.(22a) define the beta-function

$$\beta(g) \stackrel{\text{def}}{=} \left. \frac{\partial}{\partial \xi} \bar{g}_{\text{appr}}(\xi, g) \right|_{\xi=1}. \quad (42)$$

2. Integrate eq.(4), i.e., construct the function

$$f(g) \stackrel{\text{def}}{=} \int^g \frac{d\gamma}{\beta(\gamma)}, \quad (43)$$

3. Resolve the eq.(38)

$$\bar{g}_{\text{RG}}(x, g) = f^{-1}\{f(g) + \ln x\}. \quad (44)$$

4. Then, the solution  $\bar{g}_{\text{RG}}$ , precisely satisfies the RG symmetry, i.e., it is an exact solution of eq.(5) and corresponds to  $\bar{g}_{\text{appr}}$ .

For illustration, take as a  $\bar{g}_{\text{appr}}$  the simplest perturbative expression (29) for the invariant coupling. Here, the  $\beta$ -function is  $\beta(g) = -\beta_1 g^2$ , and the integration yields

$$\int_g^{\bar{g}} \frac{dg}{\beta(g)} = \frac{1}{\beta_1} \left( \frac{1}{\bar{g}} - \frac{1}{g} \right) = \ln x.$$

The solution obtained

$$\bar{g}(x, g) = \frac{g}{1 + g\beta_1 \ln x}, \quad (45)$$

on one hand, exactly satisfies the RG symmetry and, on the other, being expanded in powers of  $g$ , correlates with the input (29).

### 2.4.2 RGM usage in QFT

As it has been explained above in Section 2.1, the QFT perturbation expression of finite order does not obey the RG symmetry. On the other hand, in Section 2.4.1 it was shown that the one-loop UV approximation for  $\bar{g}$  used as an input in eq.(22) for the construction of a group generator  $\beta(g)$  yields expression (45) that obeys the group symmetry and exactly satisfies FE (5).

Now, using the geometric progression (45) as a hint, let us represent the 2-loop perturbative approximation for  $\bar{g}$  in the form

$$\bar{g}_{pt} = g - g^2 \beta_1 \ln x + g^3 [\beta_1^2 \ln^2 x - \beta_2 \ln x] + O(g^4) ,$$

where  $\beta_1$  and  $\beta_2$  mean the  $\beta$ -function coefficients at the one-loop and two-loop level, respectively. If we substitute this expression into eq.(5) we obtain

$$\bar{g}_{pt}^{(2)}(x, g) - \bar{g}_{pt}^{(2)}(x/t; \bar{g}_{pt}^{(2)}(t, g)) = g^4 \beta_1^3 \ln(x/t) \ln^2 t .$$

Meanwhile, we can use  $\bar{g}_{pt}^{(2)}$  as an input in Eq.(42). Now the step 1 yields

$$\beta^{(2)}(g) = -\beta_1 g^2 - \beta_2 g^3$$

and then (step 2)

$$\beta_1 f^{(2)}(z) = - \int^z \frac{d\gamma}{\gamma^2 + b\gamma^3} = \frac{1}{z} + b \ln \frac{z}{1+bz} ; \quad b = \frac{\beta_2}{\beta_1} . \quad (46)$$

To make the last step, we have to start with the equation

$$f^{(2)}[\bar{g}_{rg}^{(2)}(x, g)] = f^{(2)}(g) + \beta_1 \ln x$$

which is a transcendental one and has no simple explicit solution<sup>3</sup>. Due to this, one has to resolve this relation approximately. Take into account that the second, logarithmic, contribution to  $f^{(2)}(z)$  in (46) is a small correction to the first one at  $bz \ll 1$ . Under this reservation we can substitute the one-loop RG expression (45) instead of  $\bar{g}_{rg}^{(2)}$  into this correction and obtain the explicit expression

$$\bar{g}_{rg}^{(2)} = \frac{g}{1 + g\beta_1 l + g(\beta_2/\beta_1) \ln [1 + g\beta_1 l]} ; \quad l = \ln x. \quad (47)$$

This result (first obtained [8] in mid-50s) is interesting in several aspects.

First, being expanded in  $g$  and  $gl$  powers, it produces an infinite series containing “leading”, i.e.  $\sim g(gl)^n$ , and “next-to-leading”  $\sim g^2(gl)^n$  UV logarithmic contributions. Second, it contains a nontrivial analytic dependence

$$\ln(1 + g\beta_1 l) \sim \ln(\ln Q^2)$$

which is absent in the perturbation input. Third, being compared with eq.(45), it demonstrates algorithm of subsequent improving of accuracy, i.e., of RGM regularity.

Now we can resume the RGM properties. The RGM is a regular procedure of *combining* dynamical information (taken from an approximate solution) with the RG symmetry. The essence of RGM is the following:

- 1) The mathematical tool used in RGM is Lie differential equations.
- 2) The key element of RGM is possibility of (approximate) determination of group generators from dynamics.
- 3) The RGM works effectively in the case when a solution has a singular behaviour. It restores the structure of singularity compatible with RG symmetry.

<sup>3</sup>It can be expressed in terms of a special, Lambert,  $W$ -function :  $W(z) \exp^{W(z)} = z$ ; see, e.g., [19].

### 3 RG in QFT

This section is devoted, mainly, to general topics of RG applications in the QFT short distance asymptotic behavior. We discuss the specific features of UV analysis connected with use of perturbation theory, in particular, reliability of results.

#### 3.1 UV analysis in general

##### 3.1.1 One-coupling case

General analysis of the UV asymptotic behavior for the one-coupling QFT model can be performed rather simply on the basis of the solution

$$\int_g^{\bar{g}(x,g)} \frac{d\gamma}{\beta(\gamma)} = \ln x, \quad x = \frac{Q^2}{\mu^2}, \quad (48)$$

of the massless RG equation (4) for an effective coupling with  $g = g(1, g)$  and  $\mu$ , a reference point. As follows from it, the asymptotics at  $\ln x \rightarrow \infty$  corresponds to the divergence at the upper limit of the l.h.s. integral. Depending on the feature of the  $\beta$ -function the resultant UV behaviour of the invariant coupling  $\bar{g}$  differs very much.

Suppose that at very small  $g$  values the beta-function is positive. Then, three cases are possible

a) Consider first the situation with

$$\int_g^{g_*} \frac{dz}{\beta(z)} = \infty \quad (52a)$$

corresponding to the case when the beta function has a zero at some finite point  $g_*$ .

Here, the UV asymptotic value of effective coupling is finite  $g(\infty, g) = g_* < \infty$ , which relates to the *finite renormalization of the coupling constant*:  $Z = g(\infty, g)/g$ .

Using the terminology of DEs qualitative theory, one can say that at  $g = g_*$  we have a UV fixed point.

If at  $g = g_*$  there is a first order zero  $\beta(g) \simeq b(g_* - g)$ , then eq.(52a) gives

$$\bar{g}(x, g) - g_* \simeq C \exp^{-b \ln x} = C(Q^2/\mu^2)^{-b} \quad \text{as} \quad Q^2 \rightarrow \infty, \quad (50)$$

i.e., in the vicinity of a fixed point we have an asymptotic *power* regime.

b) If  $\beta(g)$  is (monotonically) increasing as  $g \rightarrow \infty$  but gentler than  $g^2$ , so that

$$\int_g^\infty \frac{dz}{\beta(z)} = \infty \quad (52b)$$

then the effective coupling tends to infinity

$$\lim_{x \rightarrow \infty} \bar{g}(x, g) \rightarrow \infty,$$

which corresponds to infinite coupling constant renormalization. Formally, this is equivalent to  $g_\infty = \infty$ .

c) At

$$\int_g^\infty \frac{d\gamma}{\beta(\gamma)} = L = \ln x_\infty < \infty, \quad (52c)$$

that happens if

$$\lim_{x \rightarrow \infty} \beta(g)/g^2 \geq \text{const},$$

the theory has an inner contradiction, as far as

$$\bar{g}(x_\infty, g) = \infty \quad \text{at} \quad x_\infty < \infty$$

and the momentum region  $x > x_\infty$  can not be described by the theory. We encounter here a *ghost trouble*, as explained below in this Section.

Up to now we have assumed that the generator  $\beta(g)$  is positive. In the opposite case

d)  $\beta(g) = -b(g) < 0$  one has to deal with the equation

$$\int_{\bar{g}}^g \frac{dz}{b(z)} = \ln x \quad (52d)$$

and study possible divergence of the integral involved at the lower limit.

If this occurs at some finite value  $g = g_\infty$ , the situation is quite analogous to the case b). The only difference is that now the effective coupling tends to its limiting value  $g_\infty$  from above.

As the most important case, we consider the possibility when the singularity lies at the origin  $\beta(0) = 0$  which happens in QCD. Then,  $\bar{g}$  vanishes  $\bar{g}(\infty, g) = 0$  in the UV limit which corresponds to the *asymptotic freedom* phenomenon. E.g., if we assume here that  $\beta(g) = -\beta_1 g^2$  at  $g \rightarrow 0$ , then

$$\bar{g}(x, g) \rightarrow \frac{1}{\beta_1 \ln x} \quad \text{as} \quad x \rightarrow \infty. \quad (51)$$

### 3.1.2 Multi-coupling case

For the quantum field model with several coupling constants one has to consider the system of coupled functional (16) or differential equations. The last ones can be analyzed by the well-known methods of the qualitative theory of differential equations.

Take the case with two coupling constants  $g$  and  $h$ . The system of evolutionary differential equations is

$$\dot{\bar{g}} = \beta_g(\bar{g}, \bar{h}), \quad \dot{\bar{h}} = \beta_h(\bar{g}, \bar{h}); \quad \dot{f} \equiv \partial f / \partial \ell; \quad \ell = \ln x. \quad (52)$$

According to (41), the general solution to this system is of the form

$$F(\bar{g}, \bar{h}) = F(g, h) + \ell, \quad \Phi(\bar{g}, \bar{h}) = \Phi(g, h) + \ell.$$

where  $F$  and  $\Phi$  — two arbitrary reversible functions.

As far as argument  $\ell = \ln x$  does not enter explicitly into the generators  $\beta_g$  and  $\beta_h$ , it can formally be excluded by dividing one of the equations (52) by the other :

$$\frac{d\bar{g}}{d\bar{h}} = F(\bar{g}, \bar{h}), \quad F = \frac{\beta_g}{\beta_h}. \quad (53)$$

This equation can be analyzed on the two dimensional phase plane  $(\bar{g}, \bar{h})$ .

First explicit example of such phase portrait has been obtained in mid-fifties by I. Ginzburg [12] — see also Section 51.4 in the third edition of the monograph [1]. The essential features are now singular points and singular solutions. Singular points correspond to  $\beta_i = 0$  (or  $= \infty$ ). They can be of different types: a stable fixed point that is known as *attractor*, an unstable fixed point and a saddle-type point. In the vicinity of the UV attractor one can have a power scaling behavior as in eq.(50). Singular solution, separatrix, joins singular points and can also be stable or unstable. Generally, the unstable ones separate the parts of phase plane with different UV asymptotes that correspond to UV stable separatrices.

## 3.2 Perturbative approach to the UV asymptote

### 3.2.1 Structure of RG results

Consider a general situation with the RG approach to the UV asymptotic behavior based on perturbation calculation input. In the one-coupling QFT case, group generators entering into DEs can be written as

$$\beta(g) = \beta_1 g^2 + \beta_2 g^3 + \dots, \quad \psi(g) = \psi_1 g + \psi_2 g^2 + \dots \quad (54)$$

Generally, expansion coefficients depend on the mass variable

$$\beta(y, g) = \sum_{l \geq 1} \beta_l(y) g^{l+1}, \quad \psi(y, g) = \sum_l \psi_l(y) g^l. \quad (55)$$

Note that if  $g$  is just the  $S$ -matrix expansion parameter (that can be equal to the coupling constant or to its square) then usually the first term in expansion for  $\beta$  is quadratic and for  $\psi$  linear, as it is explicitly indicated above.

Substituting (54) into (48), and re-expanding the ratio  $\gamma^2/\beta(\gamma)$ , we obtain after integration

$$1/g - 1/\bar{g} - \frac{\beta_1}{\beta_2} \ln(\bar{g}/g) - b_3(\bar{g} - g) + O(g^2, \bar{g}^2) = \beta_1 \ln x, \quad (56)$$

$$\ln s(x, g) = (\psi_1/\beta_1) \ln\{\bar{g}(x, g)/g\} + c_2(\bar{g} - g) + O(g^2), \quad (57)$$

$$b_3 = \beta_3/\beta_1 - (\beta_2/\beta_1)^2, \quad c_2 = [\psi_2/\beta_1](\psi_2/\psi_1 - \beta_2/\beta_1).$$

As follows, the solutions  $\bar{g}$  and  $s$  depend on two arguments  $g$  and  $g \ln x$ . By expanding them in powers of  $g$  we get

$$\bar{g}(x, g) = g f_1(g \ln x) + g^2 f_2(g \ln x) + \dots, \quad \ln s(x, g) = \varphi_1(g \ln x) + g \varphi_2(g \ln x) + \dots \quad (58)$$

where  $f_j$  and  $\varphi_i$  have a simple form. For example,  $f_1(z) = (1 - \beta_1 z)^{-1}$ ,  $\varphi_1(z) \sim f_2(z) \sim \ln f_1(z)$ .

Comparing expressions obtained with usual perturbative expansions

$$\bar{g}_{pt}(x, g) = g + g^2 \beta_1 \ln x + g^3 [\beta_1^2 \ln x + \beta_2 \ln x] + O(g^4),$$

$$s_{pt}(x, g) = 1 + g \psi_1 \ln x + g^2 [(\beta_1 \psi_1/2) \ln^2 x + \psi_2 \ln x] + O(g^3),$$

used as an input to obtain our starting generators, one can see the qualitative effect of the RGM using. In the case considered, it changes the region of applicability of the perturbation method limited by the condition  $g \ln x \ll 1$  to a more larger region defined by two relations

$$g \ll 1, \quad \bar{g}(x, g) \ll 1, \quad (59)$$

the second of which is defining.

### 3.2.2 The ghost-pole trouble

Turn now to the one-loop RG approximation for the effective coupling  $g$ , considered in the UV, i.e., massless limit. According to (45) and (23), it has the form

$$\bar{g}_{(1)}(x, g) = \frac{g}{1 - \beta_1 g \ln x}.$$

Let the numerical coefficient  $\beta_1$  be positive. Such is the case in QED where  $\beta_1 = 1/3\pi$  and  $g$  stands for the expansion parameter  $\alpha = e^2$ . This expression obviously has a pole singularity at

$$x = x_* \equiv \exp(1/\beta_1 g) = \exp(3\pi/\alpha).$$

As far as the QED effective coupling is proportional to the (transverse part of) photon propagator, this pole, generally, describes some bound state of a system with the photon quantum numbers. However, a pole related to a physical bound state must have positive residue while the l.h.s. of eq.(23) has a negative one.

This means that it corresponds not to a physical but rather to some unphysical, so-called *ghost*, state. The presence of a ghost singularity can be treated as a signal of inconsistency of a theory. Such claims have been made [20] in the mid 50s when the ghost-pole trouble was first discovered [21] just before the birth of the RGM.

The RG method proved to be very effective for a general discussion of the ghost-pole issue. The first question that must be answered here is the stability of indication of the ghost-pole existence with respect to the multi-loop corrections.

Note that in a perturbation calculation, the  $\beta$ -function depends on the adopted renormalization procedure; at the massless case, starting with the 3-loop level the coefficients of the perturbation series (54) depend on the renormalization scheme (RS) used. In QED, the 3-loop  $\beta$  function in MOM (i.e., momentum subtraction) scheme is

$$\beta_{(3)}^{MOM} = \frac{\alpha^2}{3\pi} + \frac{\alpha^3}{4\pi^2} + \frac{\alpha^4}{8\pi^3} \left( \frac{8}{3}\zeta(3) - \frac{101}{36} \right). \quad (60)$$

The numerical value of the last parenthesis is about 0.4. Neglecting it for the moment, we start our discussion with the two-loop approximation for the  $\beta$  function. According to eq.(24), the 2-loop iterative RG solution is

$$\bar{\alpha}_{(2)} = \frac{\alpha}{1 - \frac{\alpha}{3\pi}l + \frac{3\alpha}{4\pi} \ln(1 - \frac{\alpha}{3\pi}l)}.$$

This solution has an error of an order of  $\alpha^4 l$  and is interesting from several points of view. As it has been mentioned before, its  $\alpha$  expansion besides leading logs contains an infinite number of next-to-leading terms  $\alpha^2(\alpha l)^m$ , the first of which has been used as an input for construction of the  $\beta$  function. Second, in the vicinity of the ghost pole of the one-loop RG solution at  $l_1 = 3\pi/\alpha$ , the two-loop  $\bar{\alpha}_{(2)}$  solution differs from  $\bar{\alpha}_{(1)}$  considerably. Hence, an infinite sum of the next-to-leading logarithmic contributions in the region  $\alpha l \sim 1$  becomes important.

It is not trivial because for an each order of the perturbation input the next-to-leading term is negligible comparing with the leading one of the previous order (the ratio being of an order of  $\alpha\beta_2/\beta_1 = 3\alpha/4\pi \approx 2.10^{-3}$ ). It can be seen with the help of (56) that the allowing for the last, 3-loop, term in (60) also becomes essential for the  $\bar{\alpha} \sim 1$  case.

This means that the problem of existence for the ghost pole in QED cannot be solved by taking into the account of next-to-leading and so on logs. Moreover, one can argue [22] on general RG ground that it is impossible to make any qualitative statement about the UV asymptote for the  $\beta(g) > g$  case basing on RG-improved perturbation calculations. Our next example illustrates this thesis.

### 3.2.3 Scalar quartic model

For the nonlinear scalar field with the quartic (self)interaction Lagrangian

$$\mathcal{L}_{int} = -\frac{4\pi^2}{3}g\phi^4$$

important progress has been achieved in 80s in the higher perturbation orders calculation. The  $\beta$  function was calculated in the MS-scheme up to the 5-loop level [23]

$$\beta_{(5)}^{MS} = \frac{3}{2}g^2 - \frac{17}{6}g^3 + 16.275g^4 - 135.8g^5 + 1437g^6.$$

We see from this expression that, in contrast with the QED case, due to its alternate-sign structure, there is no stability here even on a qualitative level. The odd-order approximations have a

ghost-pole type behavior, whereas the even ones yield the fixed point (finite charge renormalization) case. Note also that, as can be shown [24], the upper boundary of the 10% confidence region corresponds to  $g$  values close to 0.1.

To comprehend the “loop dependence” of this boundary it is useful to represent expression  $\beta_{(5)}^{MS}$  in a slightly different form

$$\beta^{MS}(g) = \frac{3}{2}g^2 \left[ 1 - \frac{g}{0.529} + \left( \frac{g}{0.303} \right)^2 - \left( \frac{g}{0.222} \right)^3 + \left( \frac{g}{0.180} \right)^4 \right]. \quad (61)$$

It is clear now, that a boundary of the confidence region diminishes with rising the order of a loop approximation. The expression (61) looks like a beginning of a power asymptotic series of the Poincaré type. Indeed, if we represent the  $\beta$ -function in a series expansion form (54) then, as it is can be shown, the coefficients  $\beta_n$  at  $n \gg 1$  behave like  $\sim n!$ .

The method of determining asymptotic estimates of the perturbation expansion coefficients of the Green functions uses a representation in the form of a functional (i.e. path) integral. This integral written down for the mentioned expansion coefficient can be calculated by the steepest descent method in the function space. To the saddle point, there corresponds an “instanton” type Euclidean classical solution with a finite action.

In this manner, an asymptotic expression was obtained [25] for the coefficients of the  $\beta$  function expansion. It has the form

$$\beta_n \simeq \frac{1.096}{16\pi^2} n! n^{7/2} (1 + O(n^{-1})) \quad (n \rightarrow \infty). \quad (62)$$

The factorial growth of coefficients indicates that this is a power asymptotic series with zero radius of convergence that cannot be summed in the usual manner. We can obtain the information about the singularity structure at the origin ( $g = 0$ ) by using some special procedures. One of them is the Borel summation method.

Here, we give short exposition of the results on the attempt of the summation of the series of (61) type made in [24] (see also the review [26]).

Authors of the Ref. [24] used as an input the  $\beta$ -function 4-loop expression in the symmetric MOM-scheme

$$\beta_5^{MOM}(g) = \frac{3}{2}g^2 - \frac{17}{6}g^3 + 19.3g^4 - 146g^5. \quad (63)$$

This alternate-sign asymptotic series can be summed by the Borel method. The idea is to represent the sum in a form of a Laplace transform integral. It is not difficult to see that the transition to the Laplace image just “kills” the factorial factor  $n!$ . For the modified Borel transformation

$$\beta(g) = \int_0^\infty \frac{dx}{g} \exp(-x/g) \left( x \frac{\partial}{\partial x} \right)^5 B(x) \quad (64)$$

perturbation series can be written down as

$$B(x) = \sum_n \frac{\beta_n}{n!n^5} x^n.$$

It has a nonzero circle of convergence and can be summed within the circle. However, as the integration domain in (64) goes outside the convergency region, we must make an analytic continuation for the function  $B(x)$ . It can be done by a conformal transformation of the  $x$ -plane into the  $w$ -plane to map the domain of integration  $[0, \infty]$  into the interior of the unit disk and the cut  $[-\infty, -1]$  into the boundary of the disk. One can choose this transformation in such a way that it correctly reproduces the singularity on the cut. The result of conformal transformation  $x \rightarrow w(x) = (\sqrt{1+x} - 1) \cdot (\sqrt{1+x} + 1)^{-1}$  “looks quite well”:

$$B(x) = \frac{3x^2}{128} (1 - 0.32w - 0.127w^2 + 0.084w^3).$$

Then, by transformation reverse to (64) one can reconstruct beta function  $\tilde{\beta}(g)$  which is nonanalytic in the  $g$  variable with essential singularity at  $g = 0$ . Graphs of the function  $\tilde{\beta}(g)$  obtained by the Borel summation with allowance for the 1-, 2-, 3- and 4-loop approximations look very similar one to other.

They all lie now in a narrow parabolic ray slightly below the original one-loop parabola and within the limit of 10% accuracy enable to advance into the region  $g \sim 50$ . This means that the summation procedure adopted enlarges the confidence interval in several hundred times! Besides this it gives the qualitative stability of results. All they are now in favour of a ghost-type UV asymptote.

Nevertheless, these results can be considered only as a support but not the proof of the  $\phi^4$ -model inconsistency. The weak point here is that starting with (64) we have assumed the definite analyticity properties of  $\beta(g)$  in the whole complex  $g$ -plane.

### 3.3 Mass-dependent analytic solution

A general method of an approximate solution to the *massive* (i.e., mass-dependent) RG equations was developed in Ref.[27]. Analytic expressions of a high level of accuracy for an effective coupling and a one-argument function were obtained up to 3- and 4-loop order [28].

For example, the two-loop massive RG-solution for the invariant coupling

$$\alpha_s(Q^2)_{\text{rg}}^{(2)} = \frac{\alpha_s}{1 + \alpha_s A_1(Q^2, m^2) + \alpha_s A_2/A_1 \ln(1 + \alpha_s A_1(\dots))} \quad (65)$$

at small  $\alpha_s$  values corresponds to perturbation expansion

$$\alpha_s(Q^2)_{\text{pert}}^{(2)} = \alpha_s - \alpha_s^2 A_1(Q^2, m^2) + \alpha_s^3 [A_1^2 - A_2(Q^2, m^2)] + \dots$$

At the same time, it smoothly interpolates between two massless limits (with  $A_\ell \simeq \beta_\ell \ln Q^2 + c_\ell$ ) at  $Q^2 \ll m^2$  and  $Q^2 \gg m^2$  described by an equation analogous to (47). In the latter case it can be represented in the form usual for the QCD practice:

$$\bar{\alpha}_s^{-1}(Q^2/\Lambda^2)_{\text{rg}}^{(2)} \simeq \beta_1 \left\{ \ln \frac{Q^2}{\Lambda^2} + b_1 \ln \left( \ln \frac{Q^2}{\Lambda^2} \right) \right\}; \quad b_1 = \frac{\beta_2}{\beta_1^2}.$$

Solution (65) demonstrates, in particular, that the threshold crossing generally changes the subtraction scheme [29].

The investigation [27, 28] was prompted by the problem of taking explicitly into account of heavy quark masses in QCD. However, the results obtained are important from a more general point of view for a discussion of the scheme dependence problem in QFT. The method used could also be of interest for RG applications in other fields within the situation with disturbed homogeneity, such as, e.g., intermediate asymptotics in hydrodynamics, finite-size scaling in critical phenomena and the excluded volume problem in polymer theory.

In paper [30], this method was applied to the evolution of effective gauge couplings in Standard Model (SM). Here, a new analytic solution of a coupled system of three mass-dependent two-loop RG equations for three SM gauge couplings was obtained.

For this goal, one has to start with a perturbative input for the SM couplings

$$\alpha_i(Q^2, m^2) \simeq \alpha_i - \alpha_i^2 A_i(Q, m, \mu) + \alpha_i^3 A_i^2(Q) - \alpha_i^2 \sum_j \alpha_j A_{ij}(Q); \quad i = 1, 2, 3. \quad (66)$$

where  $A_i$  and  $A_{ij}$  are one- and two-loop mass- and  $\mu$ -dependent contributions of appropriate Feynman diagrams. In the framework of a massive renorm-group formalism [7, 8, 10] the corresponding Lie equations look like

$$\dot{\alpha}_i(Q^2, m^2) \simeq -\alpha_i^2(Q) \left[ \dot{A}_i(Q) + \sum_j \alpha_j(Q) \dot{A}_{ij}(Q) \right] \quad (67)$$



with  $\dot{A} \equiv \partial A / \partial \ell$ ;  $\ell = \ln Q^2 / \mu^2$ . Note that in the UV limit  $A_i(Q) = \beta_i \ell$ ;  $A_{ij}(Q) = \beta_{ij} \ell$  we arrive at the system

$$\dot{\alpha}_i(\ell) = - \left( \beta_i + \sum_j \beta_{ij} \alpha_j(\ell) \right) \alpha_i^2(\ell)$$

that is commonly used – see, e.g., Refs.[31] – for the discussion of data extrapolation across the gauge desert and possibility of Grand Unification.

The latter system can be solved iteratively in the form

$$\frac{1}{\alpha_i(\ell)} = \frac{1}{\alpha_i} + \beta_i \ell + \sum_j \frac{\beta_{ij}}{\beta_j} \ln[1 + \alpha_j \beta_j \ell]; \quad \alpha_i = \alpha_i(\mu). \quad (68)$$

Here, we present a generalization of this solution for the massive case, that is convenient for taking into account of threshold effects and discussing, in particular, the issue of the Grand Unification consistency check.

Using the method of the paper [32], one can obtain explicit iterative solution to the system (67). Here, as in the massless case, one first solves the one-loop approximation to (67) to get <sup>4</sup>

$$\alpha_i^{(1)}(Q^2, m^2) = [1/\alpha_i + A_i(Q^2, m^2)]^{-1}.$$

Inserting then this explicit expression into the second factor in the r.h.s. of (67) and performing an approximate integration of some integral – for detail see paper [32] – we arrive at the expression

$$\frac{1}{\alpha_i(Q^2, m^2)} = \frac{1}{\alpha_i} + A_i(Q) + \sum_j \frac{A_{ij}(Q)}{A_j(Q)} \ln[1 + \alpha_j A_j(Q)]. \quad (69)$$

quite analogous to (65).

The remarkable feature of this solution is that it depends explicitly only on mass-dependent perturbation coefficients  $A_i(Q)$ ,  $A_{ij}(Q)$  and, being expanded in powers of coupling constants, exactly corresponds to the perturbative input (66). On the other hand, in the massless limit it goes to solution (68).

The accuracy of the last approximate expression can be estimated by the method used in paper [34]. Generally, it corresponds to the accuracy <sup>5</sup> of three-loop expression ( $\simeq \alpha^5 \ln$ ) for the effective coupling in the one-coupling case that is quite sufficient for current practice.

### 3.4 Some important results

In the early 70s, S. Weinberg [35] proposed the notion of a *running mass* of a fermion. If considered from the viewpoint of paper [13], this idea can be formulated as follows:

*any parameter of the Lagrangian can be treated as a (generalized) coupling constant, and its effective counterpart should be included into the renorm-group formalism.*

New possibilities for applying the RG method were discovered when the technique of operator expansion at short distances (on the light cone) appeared [36]. The plausibility of this approach stems from the fact that the RG transformation, regarded as a Dyson transformation of the renormalized vertex function, involves the simultaneous scaling of all its invariant arguments – normally, the squares of the momenta. Meanwhile, for the physical amplitude, some of them are fixed on a mass shell. The expansion on the light cone, so to say, “separates the arguments”, as a result of which it becomes possible to study the physical UV asymptotic behaviour by means of the expansion coefficients (when some momenta being fixed on mass shell). As an important example, we can mention the evolution equations for moments of QCD structure functions [37].

The revealing of an *asymptotic freedom phenomenon* can be considered as the most important result obtained in particle physics by the RG technique.

<sup>4</sup>This exact solution of an one-loop massive RG equation was first obtained in Ref.[33].

<sup>5</sup>See eqs.(13) and (16) in Ref.[34].

Historically, this discovery was made [38] in the framework of the SU(3) non-Abelian Yang-Mills model in the early 70s. Since that time this model for the eight-component 4-vector field  $B_\mu^{ab}(x)$  was adopted as a basic ingredient for the QFT description of matter on the parton level.

The key point is that self-interaction of this non-abelian gluonic quantum field due to dominance of its unphysical components gives negative contribution to the beta function perturbation expansion. For the two-loop (scheme-independent) case

$$\beta_{\text{QCD}}(\alpha) = -\beta_1\alpha^2 - \beta_2\alpha^3$$

with positive  $\beta_{1,2}$  for a number  $n_f$  of quark flavours small enough.

Correspondingly, the one-loop renorm-group expression

$$\bar{\alpha}_s^{(1)}(x; \alpha_s) = \frac{\alpha_s}{1 + \alpha_s \beta_1 \ln x},$$

for the QCD effective coupling exhibits a remarkable UV asymptotic behaviour thanks to  $\beta_1$  being positive. This expression implies, in contrast to eq.(23), that the effective QCD coupling decreases as  $x \sim Q^2$  increases and tends to zero in the UV limit. This feature discovered in the early 70s, precisely corresponded to the parton physical picture of the hadronic structure.

One more interesting application of the RG method in the multicoupling case, ascending to 50s [12], refers to special solutions, the so-called separatrices in a phase space of several invariant couplings. These solutions relate effective couplings and represent scale-invariant trajectories, like, e.g.,  $g_i = g_i(g_1)$  in the phase space which are straight lines in the one-loop case.

Some of them that are “attractive” (or stable) in the UV limit, are related to symmetries that reveal themselves in the high-energy domain. It was conjectured that these trajectories may be related to *hidden symmetries of a Lagrangian* and even could serve as a tool to find them. On this basis the method was developed [39] for finding out these symmetries. It was shown that in the phase space of invariant couplings the internal symmetry corresponds to a singular solution that remains a straight-line when higher order corrections are taken into account. Such solutions corresponding to supersymmetry were derived for some combinations of gauge, Yukawa and quartic interactions.

Generally, these singular solutions obey the relations

$$\frac{dg_i}{dl} = \frac{dg_i}{dg_1} \frac{dg_1}{dl}, \quad l = \ln x$$

which are known since Zimmermann’s paper [40] as *the reduction equations*. In the 80s they were used [41] (see also review paper [42] and references therein) in the UV analyzis of asymptotically free models. Just for these cases the one-loop reduction relations are adequate to physics.

Quite recently some other application of this technique was obtained in supersymmetric generalizations of Grand Unification scenario in the Standard Model. It was shown [43, 44] that it is possible to achieve complete UV finiteness of a theory if Yukawa couplings are related to the gauge ones in a way corresponding to these special solutions, that is, to reduction relations.

The mass-dependent technique described in Section 3.3 was successfully used for the development of the Dhar-Gupta approach [45, 46] that led to finite perturbative predictions for a physical quantity which is free of renormalization scheme ambiguities. In paper [47], this approach was reformulated for the mass-dependent case with several coupling constants.

One more recent QFT development relevant to the renorm-group is the “Analytic approach” to perturbative QCD (pQCD). It is based upon the procedure of *Invariant Analyticization* [48, 49] ascending to the end of 50s.

The approach consists in the combining of two ideas: the RG summation of UV logs with analyticity in the  $Q^2$  variable, imposed by spectral representation of the Källén-Lehmann type which implements general properties of the local QFT including the Bogoliubov condition of microscopic causality. This combination was first devised [50] to get rid of the ghost pole in QED about forty years ago.

Here, the pQCD invariant coupling  $\bar{\alpha}_s(Q^2)$  is transformed into an “analytic coupling”  $\alpha_{\text{an}}(Q^2/\Lambda^2)$   $\mathcal{A}(x)$  which, by construction, is free of ghost singularities due to incorporating some nonperturbative structures.

This analytic coupling  $\mathcal{A}(x)$  has no unphysical singularities in the complex  $Q^2$ -plane; its conventional perturbative expansion precisely coincides with the usual perturbation one for  $\alpha_s(Q^2)$ ; it has no extra parameters; it obeys a universal IR limiting value  $\mathcal{A}(0) = 4\pi/\beta_0$  independent of the scale parameter  $\Lambda$ ; it turns out to be remarkably stable [49] in the IR domain with respect to higher-loop corrections and, in turn, to the scheme dependence.

Meanwhile, the “analyticized” perturbation expansion [51] for an observable  $F$ , in contrast to the usual case, may contain specific functions  $\mathcal{A}_n(x)$ , instead of powers  $(\mathcal{A}(x))^n$ . In other words the perturbation series for  $F(x)$ , due to analyticity imperative, can change its form in the IR region [19] turning into an asymptotic expansion à la Erdélyi over a nonpower set  $\{\mathcal{A}_n(x)\}$ .

## 4 RG expansion

In 70s and 80s RGM was applied to (besides QFT) critical phenomena: polymers, turbulence, non coherent radiation transfer, dynamical chaos, and so on. Simpler and less sophisticated motivation in critical phenomena (than in QFT) makes this “explosion” of RG applications possible.

### 4.1 Critical phenomena

#### 4.1.1 Spin lattices

The so called renormalization group in critical phenomena is based on the Kadanoff–Wilson procedure [52, 53] referred to as “decimation” or “blocking”. Initially, it emerged from the problem of spin lattice. Imagine a regular (two- or three- dimensional) lattice consisting of  $N^d$ ,  $d = 2, 3$  sites with an ‘elementary step’  $a$  between them. Suppose, that at every site a spin vector  $\sigma$  is sitting. The Hamiltonian describing the spin interaction of nearest neighbours

$$H = k \sum_i \sigma_i \cdot \sigma_{i+1} \quad (70)$$

contains  $k$ , the coupling constant. The statistical sum is obtained from the partition function,  $S = \langle \exp(-H/\theta) \rangle_{\text{aver.}}$ .

To realize the blocking or decimation, one has to perform the “spin averaging” over block consisting of  $n^d$  elementary sites. This is a very essential step as far as it diminishes the degree of freedom number (from  $N^d$  to  $(N/n)^d$ ). It destroys the small-range properties of the system under consideration, in the averaging course some information being lost. However, the long-range physics (like correlation length essential for phase transition) is not affected by it, and we gain simplification of our problem.

After this procedure, new effective spins  $\Sigma$  arise in sites of a new effective lattice with a step  $na$ . We obtain also a new effective Hamiltonian, with new effective coupling  $K_n$  that has to be defined in the averaging process as a function of  $k$  and  $n$

$$H_{\text{eff}} = K_n \sum_I \Sigma_I \cdot \Sigma_{I\pm 1} + \Delta H ,$$

where  $\Delta H$  contains quartic and higher terms;  $\Delta H = \sum \Sigma \cdot \Sigma \quad \Sigma \cdot \Sigma + \dots$ .

For the IR (long-distance) properties,  $\Delta H$  is unessential. Hence, we can conclude that the spin averaging leads to an approximate transformation,

$$k \sum_i \sigma \cdot \sigma \rightarrow K_n \sum_I \Sigma \cdot \Sigma , \quad (71)$$

or, taking into account the “elementary step” change, to

$$KW_n : \{a \rightarrow na, k \rightarrow K_n\} .$$

The latter is the Kadanoff-Wilson transformation.

In general, the “new” coupling constant  $K_n$  is a function of the “old” one and of the decimation index  $n$ . It is convenient to write it down in the form  $K_n = K(1/n, K)$ . Then, the KW transformation can be formulated as follows:

$$KW(n) : \left\{ a \rightarrow na, \quad k \rightarrow K_n = K\left(\frac{1}{n}, k\right) \right\} . \quad (72)$$

These transformations obey the group composition law

$$KW(n) \cdot KW(m) = KW(nm)$$

if

$$K(x, k) = K\left(\frac{x}{t}, K(t, k)\right) \quad \left[ x = \frac{1}{nm}, t = \frac{1}{n} \right] . \quad (73)$$

This is just the RG symmetry.

We observe the following points:

- The RG symmetry is approximate (due to neglecting by  $\Delta H$ ).
- The transformations  $KW(n)$  are discrete.
- There exist no reverse transformation to  $KW(n)$ .

Hence, the ‘Kadanoff-Wilson renormalization group’ is an *approximate and discrete semi-group*. For a long distance (IR limit) physics, however,  $\Delta H$  is irrelevant,  $\Delta(1/n)$  is close to continuum and it is possible to use differential Lie equations.

In application of these transformations to critical phenomena the notion of a fixed point is important. As it was explained in Section 3.1, it is usually associated with power-type asymptotic behavior. Note here that, contrary to the QFT case considered in Section 3.1, in phase transition physics we deal with the IR stable point.

#### 4.1.2 Polymer theory

In the polymer physics one considers statistical properties of polymer macromolecules which can be imagined as a very long chain of identical elements. The number of elements  $N$  could be as big as  $10^5$ , the macromolecular size reaching several hundred Angströms.

Such a big molecular chain forms a specific pattern resembling the pattern of a random walk. The central problem of the polymer theory is very close to that of a random walk and can be formulated as follows.

For a very long chain of  $N$  “steps” (the size of each step =  $a$ ) one has to find the “chain size”  $R_N$  as the distance between the “start” and the “finish” points, the distribution function of angles  $\phi_i$  between neighboring elements being given.

The function  $f(\phi)$  is defined by the forces between adjacent elements depending on some external factors like temperature  $T$ . The essential feature of a polymer chain is the impossibility of a self-intersection. This is known as an *excluded volume* effect in the random walk problem. In reality, polymer molecules are swimming in a solvent and form *globulars*.

For large  $N$  values the molecular size  $R_N$  follows the power Fleury law  $R_N \sim N^\nu$  with  $\nu$ , the Fleury index. When  $N$  is given,  $R_N$  is a functional of  $f(\phi)$  which depends on external conditions (e.g., temperature  $T$ , properties of solvent, *etc.*). If  $T$  increases,  $R_N$  increases and at some moment globulars touch one another. This is the polymerization process very similar to a phase transition phenomenon.

The Kadanov–Wilson RG (KWRG) blocking ideology has been used in polymer physics by De Gennes [54]. The key idea is a grouping of  $n$  chain subsequent elements into a new “elementary block”. This grouping operation is very close to Kadanoff’s blocking. It leads to the transformation

$$\{1 \rightarrow n ; \quad a \rightarrow A_n\}$$

which is analogous to one for spin lattice decimation. This transformation must be specified by a direct calculation which gives the explicit form of  $A_n = a(n, a)$ . Here we have a discrete semi-group. Then, by using the KWRG technique, one finds the fixed point, obtains the Fleury power law and can calculate its index  $\nu$ .

Generally, the excluded volume effect yields some complications. However, inside the QFT RG framework it can be treated rather simply [55] by introducing an additional argument similar to finite length  $L$  in transfer problem and particle mass  $m$  in QFT.

Besides polymers, the KWRG approach has been used in some fields of physics, like percolation, noncoherent radiation transfer [56], dynamical chaos [57] and some others.

Meanwhile, the original QFT-RG approach proliferated into the theory of turbulence.

#### 4.1.3 Turbulence

To formulate the turbulence problem on the “RG language” one has to perform the following steps [59, 60, 61]:

1. Introduce the generating functional for correlation functions.
2. Write the path integral representation for this functional.
3. By changing the function integration variable find the equivalence of the classical statistical system to some quantum field theory model.
4. Construct the system of Schwinger-Dyson equations for this equivalent QFT.
5. Perform the finite renormalization procedure.
6. Derive the RG equations.

## 4.2 Paths of RG expansion

RG is expanded in diverse fields of physics in two different ways:

- by direct analogy with the Kadanov-Wilson construction (averaging over some set of degrees of freedom) in polymers, non-coherent transfer and percolation, i.e., constructing a set of models for a given physical problem.
- search for an exact RG symmetry by proof of the equivalence with a QFT model: e.g., in turbulence (Refs. [59, 61]), plasma turbulence [62] and some others.

To the question Are there different renormalization groups? the answer is positive:

1. In QFT and some simple macroscopic examples (like, one dimensional transfer problem) , RG symmetry is an exact symmetry of the solution formulated in its natural variables.
2. In turbulence, continuous spin-field models and some others, it is a symmetry of an equivalent QFT model.
3. In polymers, percolation, etc. , (with KW blocking), the RG transformation is a *transformation between different auxiliary models* (specially constructed for this purpose) of a given system.

As we have shown, there is no essential difference in the mathematical formalism. There exists, however, a profound difference in physics:

— In cases 1 and 2 (as well as in some macroscopic examples), the RG is an exact symmetry of a solution.

— In the Kadanov-Wilson-type problem (spin lattice, polymers, etc. ), one has to construct a set  $\mathcal{M}$  of models  $M_i$ . The KWRG transformation

$$R(n)M_i = M_{ni} , \quad \text{with integer } n \quad (74)$$

is acting inside a set of models.

### 4.3 Two faces of RG in QFT

As it was explained in Section 1.4, the vacuum, i.e., the interparticle space, contain vacuum fluctuations. Due to them, the charge of a particle is screened. In accordance with Dirac eq.(26), in momentum space the  $Q^2$  dependence of an electron charge can be presented

$$e(Q^2) = e \left\{ 1 + \frac{\alpha}{6\pi} \ln(Q^2 r_e^2) + \dots \right\} ; \quad e^2 = e^2(1/r_e^2) = 1/137. \quad (75)$$

in terms of the classical electron charge and of electron Compton length.

The first idea of an additional symmetry in this problem was born by Stüeckelberg and Peterman [5]. In their pioneering investigation the very existence of group transformation was discovered within the renormalization procedure the result of which contains finite arbitrariness. Just this degree of freedom in finite renormalized expressions was used by Bogoliubov and Shirkov in Refs.[7]—[10]. Roughly speaking, this corresponds to the change  $r_e \rightarrow 1/\mu$ .

The basic idea was that, instead of  $1/r_e$ , one can use some other reference point  $\mu$ . This is equivalent to introducing of a *new degree of freedom* associated with the reference point scale. Instead of (75) we have

$$e(Q^2/\mu^2) = e_\mu \left\{ 1 + \frac{\alpha_\mu}{6\pi} \ln \frac{Q^2}{\mu^2} + \dots \right\} . \quad (76)$$

Here, the effective charge is considered *after* the subtracting of infinities and is given by a “finite representation” (76). The RG symmetry is formulated in terms of the  $Q^2$  scale and  $\mu$  represents the reference point.

Another approach was used by Gell-Mann and Low [6]. Their paper was devoted to the short distance behaviour in a nonlocal QED with a cutoff  $\Lambda$ , and the “ $\Lambda$  degree of freedom” was used to analyze the UV behaviour. Instead of renormalization, there is a regularization and the charge is given by the “singular representation”

$$e(\Lambda) = e \left( 1 + \frac{\alpha}{6\pi} \ln \Lambda^2 r_e^2 + \dots \right) \quad (77)$$

which is singular in the limit  $\Lambda \rightarrow \infty$ .

We can draw a transparent picture (as was commented later by Wilson in his Nobel lecture) of the last approach. Imagine an electron of a finite size, smeared over a small volume with the radius  $R_i = \hbar/c\Lambda_i$ ,  $\ln(\Lambda^2/m_e^2) \gg 1$ . The electric charge  $e_i$  of such a non-local electron is considered as depending on the cut-off momentum  $\Lambda_i$  so that this dependence accumulates the vacuum polarization effects which, in reality, take place at distances from the point electron smaller than  $R_i$ . We deal with a set of models of the non-local electron corresponding to different values of the cut-off  $\Lambda_i$ . Here,  $e_i$  depends on  $R_i$  and the vacuum polarization effects in the excluded volume  $R_i^3$  should be subtracted. In this language, the RG transformation is the transition from one value of the smearing radius to another  $R_i \rightarrow R_j$ , simultaneously with a corresponding change of the effective electron charge  $e_i \rightarrow e_j$ . In other words, the RG symmetry here is that related to operations in the space of models of non-local QED constructed in such a way that at large distances every model is equivalent to the real local one.

## 5 RG symmetry in mathematical physics

### 5.1 Functional self-similarity

The RG transformations discussed above have close connection with the concept of a self-similarity(SS). The SS transformations for problems formulated by nonlinear partial DEs are well known, since the last century, mainly in dynamics of liquids and gases. They are one parameter  $\lambda$  transformations defined as simultaneous power scaling of independent variables  $z = \{x, t, \dots\}$ , solutions  $f_k(z)$  and other functions  $V_i(z)$

$$S_\lambda : \quad \{x \rightarrow x\lambda, t \rightarrow t\lambda^a, f_k(z) \rightarrow f'_k(z') = \lambda^{\varphi_k} f_k(z'), V_i(z) \rightarrow \lambda^{\nu_i} V_i(z')\}$$

entering into the equations.

To emphasize their power structure, we use a term *power self-similarity* = PSS. According to Zel'dovich and Barenblatt, [63, 64] the PSS can be classified as:

a/ *PSS of the 1st kind*

with all indices  $a, \dots, \varphi, \nu, \dots$  being (half)integers (Integer PSS) that are usually found from the theory of dimensions;

b/ *PSS of the 2nd kind*

with irrational indices (Fractal PSS) which should be defined from dynamics.

To relate RG with PSS, let us turn to the solution of the renorm-group FE

$$\bar{g}(xt, g) = \bar{g}(x, \bar{g}(t, g)) .$$

Its general solution is known; it depends on an arbitrary function of one argument – see eq.(38). However, at the moment we are interested in a special solution linear in the second argument:  $\bar{g}(x, g) = gf(x)$ . The function  $f(x)$  should satisfy the equation  $f(xt) = f(x)f(t)$  with the solution  $f(x) = x^\nu$ . Hence,

$$\bar{g}(x, t) = gx^\nu .$$

This means that in our special case, linear in  $g$ , the RG transformation (6) is reduced to PSS transformation,

$$R_t \Rightarrow \{x \rightarrow xt^{-1}, g \rightarrow gt^\nu\} = S_t . \quad (78)$$

Generally, in RG, instead of a power law, we have arbitrary functional dependence. Thus, one can consider transformations (6), (15) and (17) as functional generalizations of usual (i.e., power) self-similarity transformations. Hence, it is natural to refer to them as to the transformations of *functional scaling* or functional (self)similarity (FS) rather than to RG-transformations. In short,

$$\text{RG} \equiv \text{FS} ,$$

with FS standing for Functional Similarity.

We can now answer the question concerning the physical meaning of the symmetry underlying FS and the Bogoliubov's renorm-group. As we have mentioned, it is not a symmetry of the physical system or the equations of the problem at hand, but a *symmetry of a solution* considered as a function of the relevant physical variables and suitable boundary conditions. A symmetry like that can be related, in particular, to the invariance of a physical quantity described by this solution with respect to the way in which the boundary conditions are imposed. The changing of this way constitutes a group operation in the sense that the group composition law is related to the transitivity property of such changes.

Homogeneity is an important feature of the physical systems under consideration. However, homogeneity can be violated in a discrete manner. Imagine that such a discrete inhomogeneity is connected with a certain value of  $x$ , say,  $x = y$ . In this case the RG transformation with the canonical parameter  $t$  will have the form (14) with the group composition law (15).

The symmetry connected with FS is a very simple and frequently encountered property of physical phenomena. It can easily be “discovered” in numerous problems of theoretical physics like classical mechanics, transfer theory, classical hydrodynamics, and so on [65, 17, 16].

## 5.2 Recent application to boundary value problem

Recently, some interesting attempts have been made to *use the RG concept in classical mathematical physics*, in particular, to study strong nonlinear regimes and to investigate asymptotic behavior of physical systems described by nonlinear PDEs.

About a decade ago, the RG ideas were applied by late Veniamin Pustovalov with co-authors [67] to analyze a problem of generating higher harmonics in plasma. This problem, after some simplification, was reduced to a couple of partial DEs with the boundary parameter – “solution characteristic” – explicitly included. It was proved that corresponding solutions admitted an exact symmetry group that takes into account transformations of this boundary parameter, which is

related to the amplitude of the magnetic field at a critical density point. The solution symmetry obtained was then used to evaluate the efficiency of harmonics generation in cold and hot plasma. The advantageous use of the RG-approach in solving the above particular problem gave promise that it may work in other cases and this was illustrated in [68] by a series of examples for various boundary value problems.

Moreover, in Refs. [65, 68] the possibility of devising a regular method for finding a special class of symmetries of solution to the boundary value problem (BVP) in mathematical physics, namely, RG-type symmetries, was discussed. The latter are defined as solution symmetries with respect to transformations involving parameters that enter through the equations as well as through the boundary conditions in addition to (or even rather than) the natural variables of the equations.

As it is well known, the aim of the modern group analysis [69, 70], which goes back to works by S. Lie [71], is to find symmetries of DEs. This approach does not include a similar problem of studying the symmetries of solutions of these equations. Outside the main direction of both the classical and modern analysis, there remains as well a study of solution symmetries with respect to transformations involving not only the variables present in the equations, but also parameters entering into the solutions from boundary conditions.

From the afore-said it is clear that the symmetries which attracted attention in the 50s in connection with the discovery of the RG in QFT were those involving the parameters of the system in the group transformations. It is natural to refer to these symmetries related to *FS* (or *RG-type*) *symmetries*.

It should be noted that the procedure of revealing the FS symmetry (FSS), or some group feature, similar to the FS regularity, in any partial case (QFT, spin lattice, polymers, turbulence and so on) up to now is not a regular one. In practice, it needs some imagination and atypical manipulation “invented” for every particular case — see the discussion in [72]. By this reason, the possibility to find a regular approach to constructing FSS is of principal interest.

Recently, a possible scheme of this kind was presented as applied to a mathematical model that is described by a BVP. The leading idea [65, 68, 73] in this case is based on the fact that solution symmetry for this system can be found in a regular manner by using the well-developed methods of modern group analysis.

The scheme that describes devising of FSS and its application is then formulated [74, 76] as follows. Firstly, a specific RG-manifold should be constructed. Secondly, some auxiliary symmetry, i.e., the most general symmetry group admitted by this manifold is to be found. Thirdly, this symmetry should be restricted on a particular solution to get the FSS. Fourthly, the FSS allows one to improve an approximate solution or, in some cases, to get an exact solution.

Depending on both a mathematical model and boundary conditions, the *first step* of this procedure can be realized in different ways. In some cases, the desired FS-manifold is obtained by including parameters, entering into a solution via an equation(s) and a boundary condition, in the list of independent variables. The extension of the space of variables involved in group transformations, e.g., by taking into account the dependence of coordinates of the renorm-group operator upon differential and/or non-local variables (which leads to the Lie—Bäcklund and non-local transformation groups [70]) can also be used for constructing the FS-manifold. The use of the Ambartsumian invariant embedding method [77] and of differential constraints sometimes allows reformulations of a boundary condition in a form of additional DE(s) and enables one to construct the FS-manifold as a combination of original and embedding equations (or differential constraints) which are compatible with these equations. At last, of particular interest is the perturbation method of constructing the FS-manifold which is based on the presence of a small parameter.

The *second step*, the calculating of a most general group  $\mathcal{G}$  admitted by the FS-manifold, is a standard procedure in the group analysis and has been described in detail in many texts and monographs — see, for example, [69, 78].

The symmetry group  $\mathcal{G}$  thus constructed cannot as yet be referred to as a renorm-group. In order to obtain this, the next, *third step* should be done which consists in restricting  $\mathcal{G}$  on a solution of a boundary value problem. This procedure utilizes the invariance condition and mathematically



appears as a “combining” of different coordinates of group generators admitted by the FS manifold.

The *final step*, i.e., constructing analytic expression for the solution of the boundary value problem on the basis of the FS, usually presents no specific problems.

A review of the results, which were obtained on the basis of the formulated scheme, can be found, for example, in [76, 79, 80].

We mention briefly, the FS analysis result for a particular problem of nonlinear optics, the problem of the laser beam self-focusing in a nonlinear medium. Here, one have a BVP for a coupled system of two nonlinear PDEs with the boundary condition given in a form of two one-argument functions. With help of RG=FS approach one new exact analytic and one new approximate analytic solution (for the practically important Gaussian initial transverse profile) has been found [81].

The important qualitative features of this example are:

- the *two-dimension singularity structure* has been analysed,
- the algebraic structure of the FSS operators is different from that of “usual RG of the QFT type”. Here, we meet with a *Lie algebra* comprising *several infinitesimal operators*.

Up to now the outlined regular method is feasible for systems that can be described by DEs and is based on the formalism of modern group analysis. However, it seems also possible to extend this approach to boundary value problems that are not described just by differential equations. A chance of such an extension is based on recent advances in group analysis of systems of integro-differential equations [82] which allow transformations of both dynamical variables and functionals of a solution to be formulated [83]. More intriguing is the issue of a possibility of constructing a regular approach for more complicated systems, in particular to those having an infinite number of degrees of freedom. The formers can be represented in a compact form by functional (or path) integrals.

## Acknowledgments

The author is grateful to Professor Ashoke Mitra for invitation to participate in this book. He is indebted to D.V. Kazakov, V.F. Kovalev, and I.L. Solovtsov for useful discussion and comments. This work was partially supported by grants of Russian Foundation for Fundamental Research (RFFR projects Nos 96-15-96030 and 99-01-00091) and by INTAS grant No 96-0842.

## References

- [1] N.N. Bogoliubov and D.V. Shirkov, *Introduction to the Theory of Quantized Fields* Wiley-Intersc., N.Y., 1959 and 1980.
- [2] N.N. Bogoliubov and D.V. Shirkov, “Problems of quantum field theory. I”, *Uspekhi Fiz. Nauk* **55**, (1955), 149-214; *ibid.* **57**, (1955), 3-92 – in Russian; *Fortschr. der Physik* **3** (1955), pp 439-495; **4** (1956), pp 438-517 – in German.
- [3] N.N. Bogoliubov and O.S. Parasyuk, *Doklady Akad. Nauk SSSR*, **100** (1955) 25–28, 429–432 – in Russian; also *Acta Mathematica*, **97** (1957), 227–266.
- [4] Laurie M. Brown, editor, *Renormalization*, Springer-Verlag, N.Y., 1993.
- [5] E.C.G. Stückelberg and A. Peterman, “La normalisation des constantes dans la theorie des quanta”. *Helv. Phys. Acta*, **26** (1953) 499-520.
- [6] M. Gell-Mann and F. Low, “Quantum Electrodynamics at Small Distances”, *Phys. Rev.* **95** (1954) 1300-1312.
- [7] N.N. Bogoliubov and D.V. Shirkov, “On the renormalization group in quantum electrodynamics”, *Doklady AN SSSR*, **103** (1955) 203-206 – in Russian.

- [8] N.N. Bogoliubov and D.V. Shirkov, "Application of the renormalization group to improve the formulae of perturbation theory", *Doklady Akad. Nauk SSSR*, **103** (1955) 391-394 – in Russian.
- [9] D.V. Shirkov, *Doklady AN SSSR*, **105** (1955) 972 – in Russian. See also in [10].
- [10] N.N. Bogoliubov and D.V. Shirkov, Charge renormalization group in quantum field theory, *Nuovo Cim.* **3** (1956) 845-637.
- [11] L.D. Landau, A.A. Abrikosov and I.M. Khalatnikov, *Doklady AN SSSR*, **95** (1954) 497; 773; 1117; **96** (1954) 261 – in Russian; *Nuovo Cim. Supp.* **3**, 80-104.
- [12] I.F. Ginzburg, *Doklady AN SSSR* **110** (1956) 535 – in Russian.
- [13] A.A. Logunov, *Soviet Phys. JETP* **3** (1956).
- [14] P.A.M. Dirac in *Theorie du Positron* (7-eme Conseil du Physique Solvay: Structure et propriete de noyaux atomiques, Oct. 1933), Gauthier-Villars, Paris, 1934, p 203.
- [15] N. Bohr, *Phys. Rev.* **48** (1935), 696.
- [16] D.V. Shirkov, *Sov. Phys. Doklady* **27** (1982) 197; The RG method and functional self-similarity in physics – in *Nonlinear and turbulent processes in physics*, Ed. R.Z. Sagdeev, Harwood Acad.Publ., N.Y. 1984, v.3, pp 1637-1647; D.V. Shirkov, "Renormalization group in different fields of theoretical physics", KEK Report 91-13, Feb. 1992.
- [17] M.A. Mnatsakanyan. *Doklady AN SSSR*, **262** (1982) 856-860. English transl. in *Soviet Phys. Doklady* **27** (1982).
- [18] L.V. Ovsyannikov, *Doklady AN SSSR*, **109** (1956) 1112. For English translation see pp 76-79 in *In the intermission ...*, Ed. Yu.A. Trutnev, WS, 1998.
- [19] D.V. Shirkov, *Theor. Math. Fiz.* **119** (1999) 55; hep-th/9810246.
- [20] L.D. Landau and I.Ya. Pomeranchuk, *Doklady AN SSSR*, **102** (1955) 489-492; I.Ya. Pomeranchuk, *ibid.*, **103** (1955) 1005; **105** (1955) 461 – in Russian; *Nuovo Cim.* **10** (1956) 1186; see also L.D. Landau, *On the Quantum Theory of Fields* – in *Niels Bohr and the development of physics*, eds. W. Pauli et al., Pergamon, London, 1955, pp 52-69.
- [21] E.S. Fradkin, *Zh. Eksp. Teor. Fiz.* **28** (1955) 750-752; English transl. in *Soviet Phys. JETP* **1** (1955).
- [22] N.N. Bogoliubov and D.V. Shirkov, *Doklady AN SSSR* **105** (1955) 685-688.
- [23] S.G. Gorishny et al., *Phys. Lett.* **132B** (1983) 351. D.I. Kazakov, *Phys. Lett.* **133B** (1983) 406.
- [24] D.I. Kazakov, O.V. Tarasov and D.V. Shirkov, *Teor. Mat. Fiz.* **38** (1979) 15.
- [25] L.N. Lipatov, *Zh. Eksp. Teor. Fiz.* **71** (1976) 2010.
- [26] D.I. Kazakov and D.V. Shirkov, *Fortschr. d. Phys.* **28** (1980) 465-499.
- [27] D.V. Shirkov, *Sov. J. Nucl. Phys.* **34(2)** (1981) 300-2; *Theor. Math. Fiz.* **49** (1981) 1039-42.
- [28] D.V. Shirkov, *Nucl. Phys.* **B 371** (1992) 467-81.
- [29] D.V. Shirkov, Mass Effects in Running Coupling Evolution and Hard Processes, in *Perspectives in Particle Physics*, Eds. D. Klabucar et al., WS, 1995, pp 1-13; On continuous mass-dependent analysis of DIS data, in *Proc. EPSHEP95 Conf. (Bruxelles, July 1995)*, Eds. J. Lemonne et al., WS, pp 141-2.

- [30] D.V. Shirkov, Mass and Scheme Effects in Coupling Constant Evolution, *Teor. Mat. Fizika* (1992) **93** 466-72.
- [31] U. Amaldi et al., *Phys. Lett.* **B260** (1991) 447-55; M.B. Einhorn and D.R.T. Jones, *Nucl. Phys.* **B 196** (1982) 475.
- [32] D.V. Shirkov, *Sov. J. Nucl. Phys.* **34** (1981) 300-2; *Teor. Mat. Fiz. (USSR)* **49** (Dec. 1981) 291-7 [pp 1039-43 in the American ed.].
- [33] V.Z. Blank and D.V. Shirkov, *Nucl. Phys.* **2** (1956) 356-70.
- [34] D.V. Shirkov, *Nucl. Phys.* **B 371** (1992) 467-81.
- [35] S. Weinberg, *Phys. Rev.* **D 8** (1973) 605-625.
- [36] K. Wilson, *Phys. Rev.* **179** (1969) 1499-1515.
- [37] G. Altarelli and G. Parisi, *Nucl. Phys.* **B 126** (1977) 298-318.
- [38] D. Gross and P. Wilczek, *Phys. Rev.* **D8** (1973) 3633-52;  
H. Politzer, *Phys. Rev. Lett.* **30** (1973) 1346-49.
- [39] D.I. Kazakov and D.V. Shirkov, Singular Solutions of RG Eqs. and the Symmetry of the Lagrangian, JINR Preprint E2-8974, 1975, in High Energy Particle Interaction (Proceed. 1975 Smolenice Conf.), Eds. D. Krupa & J. Pisut, Veda, Bratislava, 1976, 255-78.
- [40] W. Zimmermann, *Comm. Math. Phys.* **97** (1985) 211.
- [41] R. Oehme and W. Zimmermann, *Comm. Math. Phys.* **97** (1985) 569; R. Oehme, K. Sibold and W. Zimmermann, *Phys. Lett.* **B 147** (1984) 115; **B153** (1985) 142.
- [42] W. Zimmermann, in Renormalization Group, Eds. D.V. Shirkov et al, WS, Singapore, 1988, pp 55-64.
- [43] A.V. Ermushev, D.I. Kazakov, O.V. Tarasov, *Nucl. Phys.* **B 281** (1987) 72.
- [44] D.I. Kazakov, *Phys. Lett.* **B 421** (1998) 211-216.
- [45] A. Dhar, *Phys. Lett.* **128 B** (1983) 407.
- [46] A. Dhar and V.G. Gupta, *Pramana* **21** (1983) 207; *Phys. Rev.* **D 29** (1984) 2822.
- [47] V.G. Gupta, D.V. Shirkov and O. Tarasov, *Intern. J. Mod. Phys.* **A 6** (1991) 3381.
- [48] D.V. Shirkov and I.L. Solovtsov, *JINR Rapid Comm.* No. 2[76]-96 (1996) 5, hep-ph/9604363; *Phys. Rev. Lett.* **79** (1997) 1209, hep-ph/9704333.
- [49] D.V. Shirkov, *Nucl. Phys. (Proc. Suppl.)* **B 64**, (1998) 106, hep-ph/9708480.
- [50] N.N. Bogoliubov, A.A. Logunov and D.V. Shirkov, *Sov. Phys. JETP* **10** (1959) 574.
- [51] K.A. Milton, I.L. Solovtsov and O.P. Solovtsova, *Phys. Lett.* **B 415** (1997) 104.
- [52] L. Kadanoff, *Physica* **2** (1966) 263.
- [53] K. Wilson, *Phys. Rev.* **B4** (1971) 3174-3183.
- [54] P.G. De Gennes, *Phys. Lett.* **38A** (1972), 339-40; J. des Cloiseaux, *J. Physique (Paris)* **36** (1975) 281.
- [55] V.I. Alkhimov, *Theor. Mat. Fiz.* **39** (1979) 281; **59** (1984) 432.

- [56] T.L. Bell et al., *Phys. Rev.* **A17** (1978) 1049-1057;  
G.F. Chapline, *Phys. Rev.* **A21** (1980) 1263-1271.
- [57] B.V. Chirikov, *Lecture notes in physics* **179** (1983) 29; B.V. Chirikov & D.L. Shepelansky, Chaos Border and Statistical Anomalies – in [58], 221; Yu. G. Sinai & K.M. Khanin, Renormalization group method in the theory of dynamical systems – in [58], 251; A. Peterman and A. Zichichi, *Nuovo Cimento* **109A** (1996) 341.
- [58] Renormalization Group, (Proceed. 1986 Dubna Conference), Eds. D.V. Shirkov, D.I. Kazakov and A.A. Vladimirov, WS, Singapore, 1988.
- [59] C. DeDominicis and P. Martin, *Phys. Rev.* **A19** (1979) 419-422.
- [60] L. Adjemyan et al, *Teor. Mat. Fiz.* **58** (1984) 72; **65** (1985) 196; A.N. Vasiliev, Quantum Field Renormalization Group in the Theory of Turbulence and in Magnetic Hydrodynamics, in [58], pp 146-159.
- [61] A.N. Vasiliev, Quantum Field Renormalization Group in the Theory of Critical Behavior and Stochastic Dynamics, PINF Publ., St-Petersburg, 1998, 773 pp – in Russian, also Gordon & Breach, Amsterdam (in press).
- [62] G. Pelletier, *Plasma Phys.* **24** (1980) 421.
- [63] Ja.B. Zel'dovich and G.I. Barenblatt, *Sov. Phys. Doklady* **3**(1) (1958) 44-47; see also in [64].
- [64] G.I. Barenblatt, Scaling, self-similarity, and intermediate asymptotics, Cambridge Univ. Press, 1996.
- [65] D.V. Shirkov, Renormalization group in modern physics, in [58] pp 1-32, *Int. J. Mod. Phys.* **A3** (1988) 1321-1341; Several topics on renorm-group theory, in [66] , pp 1-10.
- [66] Renormalization Group '91 (Proceed. of 1991 Dubna Conf.), Eds. D.V. Shirkov and V.B. Priezzhev, WS, Singapore, 1992.
- [67] V.F. Kovalev and V.V. Pustovalov, *Teor. Mat. Fizika* **81** (1990) 1061–1071; also: Strong nonlinearity and generation of high harmonics in laser plasma, in Proceed. Conf. Plasma Physics (Kiev, USSR, April 1987), Ed. A.G. Sitenko, Naukova Dumka, Kiev, 1987, **1**, 271; Influence of laser plasma temperature on the high harmonics generation process, *ibid.*, **1**, 274.
- [68] V.F. Kovalev, S.V. Krivenko, V.V. Pustovalov, The Renormalization group method based on group analysis – in [66], 300-314.
- [69] L.V. Ovsyannikov, Group analysis of differential equations, Acad. Press, N.Y., 1982.
- [70] N.H. Ibragimov, Transformation groups applied to mathematical physics, Reidel Publ., Dordrecht-Lancaster, 1985.
- [71] M. S. Lie, *Gesammelte Abhandlungen*, Leipzig-Oslo, Bd.5, 1924; Bd.6, 1927.
- [72] D.V. Shirkov, Bogoliubov renormgroup, *Russian Math. Surveys* **49:5** (1994) 155 – with misprints. For corrected version see: The Bogoliubov Renormalization Group (second English printing), JINR Comm. E2-96-15; hep-th/9602024.
- [73] D.V. Shirkov, *Intern. J. Mod. Physics C* **6** (1995) 503-512.
- [74] V.F.Kovalev, RG-symmetries: constructing and applications – in [75], 263.
- [75] Renormalization Group '96, (Proc. 1996 Dubna Conf.), Eds. D.V. Shirkov, D.I. Kazakov and V.B. Priezzhev, JINR Publ., Dubna, 1997.

- [76] V.F. Kovalev, V.V. Pustovalov, D.V. Shirkov, *J. Math. Phys.* **39** (1998) 1170-1188; hep-th/9706056.
- [77] The embedding method was introduced by V.A.Ambartzumyan in *Astr. Journ.* **19** (1942) 30 – in Russian; later on this method enjoyed wide application to different problems – see, e.g., J.Casti, R.Kalaba, *Imbedding methods in applied mathematics* Addison-Wesley, Reading, Ma, 1973 and references therein.
- [78] Peter J. Olver, *Applications of Lie groups to differential equations*, Springer, N. Y., 1986; *CRC Handbook of Lie Group Analysis of Diff. Equations*, Ed. N.H.Ibragimov (CRC Press, Boca Raton, Florida, USA) – in three volumes. 1994 – 1996.
- [79] V.F.Kovalev, Group and renormgroup symmetries of boundary value problems, in *Modern group analysis VI*, Eds. N.H.Ibragimov and F.M.Mahomed, 1997, New Age Internat'l (P) Ltd Publ., India, N. Delhi, 225.
- [80] V.F. Kovalev and D.V. Shirkov, *Teor. Mat. Fizika* **121** (Oct. 1999) No.1.
- [81] V.F. Kovalev and D.V. Shirkov, *J. Nonlin. Opt. Phys. & Mater.* **6** (1997) 443.
- [82] V.F. Kovalev, S.V. Krivenko and V.V. Pustovalov, *Differ. Equations* **29** (1993) No 10, 1568; No 11, 1712.
- [83] V.F. Kovalev, S.V. Krivenko and V.V. Pustovalov, *J. Nonlin. Math. Phys.* **3** (1996) 175-180.

## 2. The Similarity Renormalization Group

Robert J. Perry and Sérgio Szpigel \*

Department of Physics

The Ohio State University, Columbus, OH 43210

### Abstract

Quantum field theories require a cutoff to regulate divergences that result from local interactions, and yet physical results can not depend on the value of this cutoff. The renormalization group employs a transformation that changes the cutoff to isolate hamiltonians that produce cutoff-independent eigenvalues. The similarity renormalization group is based on similarity transformations that regulate off-diagonal matrix elements, forcing the hamiltonian towards a band-diagonal form as the cutoff is lowered. This avoids pathologies that plagued traditional transformations acting on hamiltonians, making it possible to produce a well-behaved perturbative approximation of renormalized hamiltonians in asymptotically free theories. We employ a simple two-dimensional delta function example to illustrate this new renormalization technique.

## 1 Introduction

Early attempts to combine quantum mechanics and special relativity led to the consideration of local interactions, which are consistent with causality and avoid signals that propagate faster than light. Local interactions lead to divergences in perturbation theory, whose discovery caused some of the best theorists in the world to question the foundations of quantum mechanics. Eventual successes at fitting precise atomic experimental data led to the universal acceptance of renormalization recipes that were acknowledged to make little sense [1]. Initially the perturbative renormalization of QED required theorists to match perturbative expansions in powers of a bare and physical electronic charge [2], but the bare charge clearly diverges logarithmically in QED and the success of an expansion in powers of such a coupling was mysterious at best [3].

The first steps towards making sense of renormalization theory were taken in the 1950's with the invention of the perturbative renormalization group [4, 5, 6, 7, 8], although serious investigators found the theory was still plagued by non-convergent sums because QED is not asymptotically free. The development of Wilson's renormalization group formalism [9, 10] and the discovery of asymptotic freedom [11] allowed physicists to produce a logically reasonable picture of renormalization in which perturbative expansions at any high energy scale can be matched with one another, with no necessity to deal with intermediate expansions in powers of a large parameter.

In this pedagogical article we take advantage of the fact that the divergences in field theory result entirely from local interactions. To understand the most important aspects of renormalization theory requires only a background in nonrelativistic quantum mechanics, because as has been long known the divergences of field theory are directly encountered when one tries to impose locality on the Schrödinger equation. In this case the interactions we consider that are local at all scales are delta functions and derivatives of delta functions. These divergences can be regulated by the introduction of a cutoff, and the artificial effects of this cutoff must be removed by renormalization. The simplicity of the one-body Schrödinger equation makes it possible to renormalize the theory exactly, disentangling the effects of locality from the complicated many-body effects and symmetry constraints encountered in realistic field theories. There is a large literature on the subject [12]-[29], primarily pedagogical.

---

\*E.mail:perry@mps.ohio-state.edu

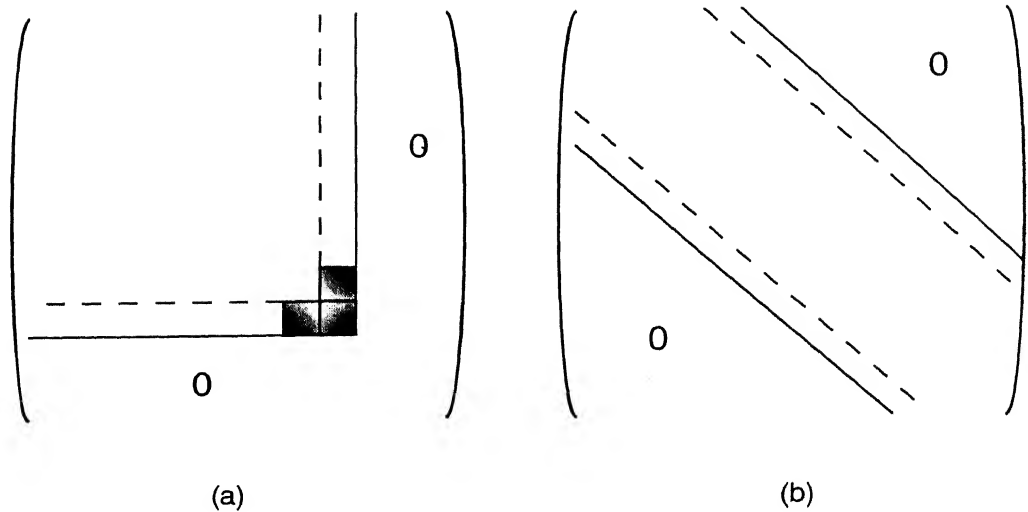


Figure 1: Two ways to run a cutoff on free energy. In (a) a cutoff on the magnitude of the energy is lowered from the solid to the dashed lines, with problems resulting from the removed shaded region. In (b) a cutoff on how far off diagonal matrix elements appear is lowered from the dashed to the solid lines.

The similarity renormalization group (SRG) is a very recent development invented by Stan Głazek and Ken Wilson [30, 31], and independently by Franz Wegner [32]. We do not review the applications of this method, which are growing in number.

In the SRG, as in Wilson's original renormalization group formalism [33, 34], transformations that explicitly run the cutoff are developed. These transformations are the group elements that give the renormalization group its name.

In his earliest work [33, 34] Wilson exploited a transformation originally invented by Claude Bloch [35]. It uses a cutoff on the states themselves, and as the cutoff is lowered, states are removed from the Hilbert space. If the hamiltonian is viewed as a matrix, these cutoffs can be seen as limiting the size of this matrix and the transformation reduces this size, as illustrated in Fig. 1a. Wilson introduced a rescaling operation to allow transformed hamiltonians to be compared with initial hamiltonians, despite the fact that they act on different spaces; however, the Bloch transformation is ill-defined and even in perturbation theory it leads to artificial divergences. These divergences come from the small energy differences between states retained and states removed by the transformation, and they appear in the form of small energy denominators in the perturbative expansion of the transformed hamiltonian. These small energy denominator problems led Wilson to abandon the hamiltonian formulation of field theory in favor of path integral formulations, but the virtues of the hamiltonian formulation over the path integral formulation for many problems remains.

The breakthrough provided by the SRG is that the transformations are typically unitary, making them well-defined, and they run a cutoff on energy differences rather than on individual states, as illustrated in Fig. 1b. Again viewing the hamiltonian as a large matrix, these cutoffs limit the off-diagonal matrix elements and as they are reduced the hamiltonian is forced towards diagonal form. The perturbative expansion for transformed hamiltonians contains no small energy denominators, so the expansion breaks down only when interactions become sufficiently strong, in which case perturbation theory should fail in any case.

Although the SRG has not yet been applied to a wide range of problems, it may be an important

new tool both for attacking field theories and non-relativistic many-body problems.

When the SRG is used with coupling coherence [36, 37], which we explain below, it allows us to construct effective theories with the same number of free parameters as the underlying ‘fundamental’ theory. For the delta-function example there is one fundamental parameter, the strength of the regulated delta-function as the cutoff is removed. In the SRG with coupling coherence, there is only one fundamental coupling and all new couplings are perturbative functions of the fundamental coupling that are given by coupling coherence. It is the renormalization group flow of the added couplings, and a boundary condition that they vanish when the fundamental coupling is taken to zero, that fixes their dependence on the fundamental coupling.

The examples we use in this article do not illustrate non-Gaussian fixed points, so their scaling properties are driven by naive dimensional analysis. However, we will see that even in these cases scaling behavior of effective hamiltonians derived using a perturbative similarity renormalization group can be very complicated. We will see that in the perturbative SRG there are errors arising from the approximate treatment of the fundamental running coupling and the approximate treatment of the relation between this coupling and the new couplings of irrelevant operators.

In a realistic calculation the marginal coupling, which corresponds to the strength of the regulated delta function, would be fit to data. In order to clearly illustrate the logarithmic errors that result from using the perturbative SRG equations, we approximate this marginal coupling in this article rather than renormalizing it nonperturbatively by fitting data. The strengths of the irrelevant operators, which correspond to derivatives of the regulated delta function, are approximated using expansions in powers of the approximate running coupling that are fixed by coupling coherence. The approximate running coupling differs from the exact running coupling by inverse powers of logarithms of the cutoff, and the error analysis for the binding energy displays the resultant inverse logarithmic errors in addition to power-law errors seen in all approximate renormalization group calculations. In addition there are errors in the strengths of the irrelevant operators resulting from using a truncated expansion in powers of the running coupling and an approximate running coupling, both of which introduce inverse logarithmic errors in addition to the power-law errors normally seen.

The utility of the renormalization group rests on our ability to accurately determine and control the magnitude of errors resulting from the artificial cutoff. For perturbative calculations this issue is not critical, but in all field theories (and in our example) a scale is reached where the coupling becomes large and a non-perturbative calculation must be done. The renormalization group allows us to eliminate as much perturbative physics as possible (*i.e.*, lower the cutoff as far as possible in an asymptotically free theory), so that the essential degrees of freedom that couple non-perturbatively can be isolated.

## 2 Similarity Renormalization Group

In this section we review the general formulation of the SRG developed by Głazek and Wilson [30, 31] and a specific transformation developed by Wegner [32]. The reader may wish to skip the general formulation on a first reading.

### 2.1 Głazek-Wilson Formulation

Consider a system described by a hamiltonian written in the form

$$H = h + V, \quad (1)$$

where  $h$  is the free hamiltonian and  $V$  is an interaction.

In general, the hamiltonian can couple states of all energy scales and such couplings can be a source of ultraviolet divergences. The goal of the SRG is to obtain an effective hamiltonian in which the couplings between high and low-energy states are removed, while avoiding any problems from small energy denominators in effective interactions. The procedure is implemented by a



unitary transformation that generates effective interactions that reproduce the effects of the eliminated couplings. The effective hamiltonian cannot produce ultraviolet divergences at any order in perturbation theory as long as its matrix elements are finite.

In our discussion we will use the basis of eigenstates of the free hamiltonian,

$$h|i\rangle = \epsilon_i|i\rangle . \quad (2)$$

We start by defining a bare hamiltonian,  $H_\Lambda$ , regulated by a very large cutoff  $\Lambda$  (here with dimensions of energy) on the change in free energy at the interaction vertices,

$$H_\Lambda \equiv h + V_\Lambda , \quad (3)$$

$$V_\Lambda \equiv f_\Lambda \bar{V}_\Lambda , \quad (4)$$

$$\bar{V}_\Lambda \equiv v + H_\Lambda^{ct} , \quad (5)$$

where  $f_\Lambda$  is a “similarity function”,  $\bar{V}_\Lambda$  is defined as the reduced interaction and  $H_\Lambda^{ct}$  are counterterms that must be determined through the process of renormalization in order to remove  $\Lambda$  dependence in physical quantities.

The similarity function  $f_\Lambda$  regulates the hamiltonian by suppressing matrix elements between free states with significantly large energy difference and acts in the following way:

$$\begin{aligned} \langle i|f_\Lambda H_\Lambda|j\rangle &\equiv \epsilon_i \delta_{ij} + f_\Lambda(\epsilon_i - \epsilon_j) \langle i|\bar{V}_\Lambda|j\rangle \\ &\equiv \epsilon_i \delta_{ij} + f_{\Lambda ij} \bar{V}_{\Lambda ij} . \end{aligned} \quad (6)$$

Typically, the similarity function is chosen to be a smooth function satisfying

$$\begin{aligned} (i) f_\Lambda(\epsilon_i - \epsilon_j) &\rightarrow 1, \text{ when } |\epsilon_i - \epsilon_j| \ll \Lambda , \\ (ii) f_\Lambda(\epsilon_i - \epsilon_j) &\rightarrow 0, \text{ when } |\epsilon_i - \epsilon_j| \gg \Lambda . \end{aligned} \quad (7)$$

In several papers the similarity function has been chosen to be a step function. Although useful for doing analytic calculations, such a choice can lead to pathologies.

The similarity transformation is defined to act on the bare regulated Hamiltonian,  $H_\Lambda$ , lowering the cutoff down to a scale  $\lambda$ :

$$H_\lambda \equiv U(\lambda, \Lambda) H_\Lambda U^\dagger(\lambda, \Lambda) . \quad (8)$$

The renormalized Hamiltonian can be written in the general form

$$H_\lambda \equiv h + V_\lambda , \quad (9)$$

$$V_\lambda \equiv f_\lambda \bar{V}_\lambda . \quad (10)$$

The transformation is unitary, so  $H_\Lambda$  and  $H_\lambda$  produce the same spectra for observables. Also, if an exact transformation is implemented, the physical predictions using the renormalized Hamiltonian must be independent of the cutoff  $\lambda$  and  $H_\Lambda^{ct}$  is chosen so that they also become independent of  $\Lambda$  as  $\Lambda \rightarrow \infty$ .

The unitarity condition is given by:

$$U(\lambda, \Lambda) U^\dagger(\lambda, \Lambda) \equiv U^\dagger(\lambda, \Lambda) U(\lambda, \Lambda) \equiv 1 . \quad (11)$$

The similarity transformation  $U$  can be defined in terms of an anti-hermitian operator  $T_\lambda$  ( $T_\lambda^\dagger = -T_\lambda$ ) which generates infinitesimal changes of the cutoff energy scale,

$$U(\lambda, \Lambda) \equiv \mathcal{T} \exp \left( \int_\lambda^\Lambda T_{\lambda'} d\lambda' \right) , \quad (12)$$

where  $\mathcal{T}$  orders operators from left to right in order of *increasing* energy scale  $\lambda'$ . Using

$$T_\lambda = U(\lambda, \Lambda) \frac{dU^\dagger(\lambda, \Lambda)}{d\lambda} = -\frac{dU(\lambda, \Lambda)}{d\lambda} U^\dagger(\lambda, \Lambda) , \quad (13)$$

and the unitarity condition Eq. (11), we can write Eq. (8) in a differential form,

$$\frac{dH_\lambda}{d\lambda} = [H_\lambda, T_\lambda] . \quad (14)$$

This is a first-order differential equation, which is solved with the boundary condition  $H_\lambda|_{\lambda \rightarrow \Lambda} \equiv H_\Lambda$ . The bare Hamiltonian is typically given by the canonical Hamiltonian plus counterterms that must be uniquely fixed to complete the renormalization.

The operator  $T_\lambda$  is defined by specifying how  $\bar{V}_\lambda$  and  $h$  depend on the cutoff scale  $\lambda$ . For simplicity in this article, we demand that  $h$  is independent of  $\lambda$ , although this may not lead to an increasingly diagonal effective hamiltonian in all cases. We also demand that no small energy denominators can appear in the hamiltonian. These constraints are implemented by the conditions

$$\frac{dh}{d\lambda} \equiv 0 , \quad (15)$$

$$\frac{d\bar{V}_\lambda}{d\lambda} \equiv [V_\lambda, T_\lambda] . \quad (16)$$

To obtain the renormalized Hamiltonian perturbatively, expand

$$\bar{V}_\lambda = \bar{V}_\lambda^{(1)} + \bar{V}_\lambda^{(2)} + \dots , \quad (17)$$

$$T_\lambda = T_\lambda^{(1)} + T_\lambda^{(2)} + \dots , \quad (18)$$

$$H_\lambda^{ct} = H_\lambda^{(2),ct} + H_\lambda^{(3),ct} + \dots , \quad (19)$$

where the superscripts denote the order in the original interaction,  $V$ . A general form of these effective interactions is

$$\bar{V}_\lambda^{(i)} = - \sum_{j,k=1}^{\infty} \delta_{(j+k,i)} \int_\lambda^\Lambda d\lambda' [V_{\lambda'}^{(j)}, T_{\lambda'}^{(k)}] + H_\lambda^{(i),ct} , \quad (20)$$

for  $i = 2, 3, \dots$ , with  $\bar{V}_\lambda^{(1)} = v$ . For instance, the explicit form of the second-order effective interaction  $\bar{V}_\lambda^{(2)}$  is

$$\bar{V}_{\lambda ij}^{(2)} = \sum_k V_{ik} V_{kj} \left( \frac{g_{ikj}^{(\lambda\Lambda)}}{\Delta_{ik}} + \frac{g_{jki}^{(\lambda\Lambda)}}{\Delta_{jk}} \right) + H_{\lambda ij}^{(2),ct} , \quad (21)$$

where

$$g_{ikj}^{(\lambda\Lambda)} \equiv \int_\lambda^\Lambda d\lambda' f_{\lambda'jk} \frac{df_{\lambda'ki}}{d\lambda'} , \quad (22)$$

$$\Delta_{ij} = \epsilon_i - \epsilon_j . \quad (23)$$

The counterterms  $H_\lambda^{(n),ct}$  can be determined order-by-order using the idea of coupling-coherence [36, 37]. This is implemented by requiring the hamiltonian to reproduce itself in form under the similarity transformation, the only change being explicit dependence on the running cutoff in the operators and the implicit cutoff dependence in a finite number of independent running couplings. All other couplings depend on the cutoff only through their dependence on the independent couplings. In general, we also demand the dependent couplings to vanish when the independent couplings are taken to zero; i.e, the interactions are turned off. If the only independent coupling in the theory is  $\alpha_\lambda$ , the renormalized hamiltonian can be written as an expansion in powers of this coupling:

$$H_\lambda = h + \alpha_\lambda \mathcal{O}^{(1)} + \alpha_\lambda^2 \mathcal{O}^{(2)} + \dots . \quad (24)$$

In this way, the effective hamiltonian obtained using the similarity transformation is completely determined by the underlying theory. The procedure can be extended to arbitrarily high orders, although it becomes increasingly complex both analytically and numerically.

## 2.2 Wegner Formulation

The Wegner formulation of the SRG [32] is defined in a very elegant way in terms of a flow equation analogous to the SRG Equation in the Głazek-Wilson formalism [30, 31],

$$\frac{dH_s}{ds} = [H_s, T_s] . \quad (25)$$

Here the hamiltonian  $H_s = h + v_s$  evolves with a flow parameter  $s$  that ranges from 0 to  $\infty$ . The flow-parameter has dimensions  $1/(\text{energy})^2$  and is given in terms of the similarity cutoff  $\lambda$  by  $s = 1/\lambda^2$ .

In Wegner's scheme the similarity transformation is defined by an explicit form for the generator of the similarity transformation,  $T_s = [H_s, H_0]$ , which corresponds to the choice of a gaussian similarity function with uniform width. In the original formulation, Wegner advocates the inclusion of the full diagonal part of the hamiltonian at scale  $s$  in  $H_0$ . For a perturbative calculation of  $H_s$ , we can use the free hamiltonian,  $H_0 = h$ . With this choice, the flow equation for the hamiltonian is given by

$$\frac{dH_s}{ds} = [H_s, [H_s, h]] . \quad (26)$$

The reduced interaction,  $\bar{V}_{sij}$  (the interaction with the gaussian similarity function factored out) is defined by

$$V_{sij} = f_{sij} \bar{V}_{sij} , \quad (27)$$

$$f_{sij} = e^{-s\Delta_{ij}^2} . \quad (28)$$

Assuming that the free hamiltonian is independent of  $s$ , we obtain the flow equation for the reduced interaction,

$$\frac{d\bar{V}_{sij}}{ds} = \sum_k (\Delta_{ik} + \Delta_{jk}) \bar{V}_{sik} \bar{V}_{skj} e^{-2s\Delta_{ik}\Delta_{jk}} , \quad (29)$$

where we use  $\Delta_{ij}^2 - \Delta_{ik}^2 - \Delta_{jk}^2 = -2\Delta_{ik}\Delta_{jk}$ . We should emphasize that this is an exact equation.

To solve this equation we impose a boundary condition,  $H_s|_{s \rightarrow s_0} \equiv H_{s_0}$ . Then, we make a perturbative expansion,

$$\bar{V}_s = \bar{V}_s^{(1)} + \bar{V}_s^{(2)} + \dots , \quad (30)$$

where the superscript implies the order in the bare interaction  $\bar{V}_{s_0}$ . It is important to observe that counterterms are implicit in the bare interaction and can be determined in the renormalization process using coupling coherence.

At first order we have

$$\frac{d\bar{V}_{sij}^{(1)}}{ds} = 0 , \quad (31)$$

which implies

$$\bar{V}_{sij}^{(1)} = \bar{V}_{s_0ij} , \quad (32)$$

where  $s$  is the final scale. Because of the dimensions of the flow parameter we have  $s > s_0$ , corresponding to a smaller cutoff. The "no cutoff limit" corresponds to  $s_0 \rightarrow 0$ .

At second order we have

$$\frac{d\bar{V}_{sij}^{(2)}}{ds} = \sum_k (\Delta_{ik} + \Delta_{jk}) \bar{V}_{s_0ik} \bar{V}_{s_0kj} e^{-2s\Delta_{ik}\Delta_{jk}} . \quad (33)$$

Integrating, we obtain

$$\begin{aligned} \bar{V}_{sij}^{(2)} &= \frac{1}{2} \sum_k \bar{V}_{s_0ik} \bar{V}_{s_0kj} \left( \frac{1}{\Delta_{ik}} + \frac{1}{\Delta_{jk}} \right) \times \\ &\quad \times [e^{-2s_0\Delta_{ik}\Delta_{jk}} - e^{-2s\Delta_{ik}\Delta_{jk}}] . \end{aligned} \quad (34)$$

By construction, the Wegner transformation is unitary and avoids small energy denominators. The Wegner transformation is one of the Glazek-Wilson transformations, with the similarity function chosen to be  $f_{\lambda ij} = e^{-\Delta_{ij}^2/\lambda^2}$ .

### 2.3 Strategy

In our applications of the SRG we use Wegner's transformation. The renormalized hamiltonian for the non-relativistic delta-function potential in D-dimensions is given by

$$H_\lambda(\mathbf{p}, \mathbf{p}') = p^2 \delta^{(D)}(\mathbf{p} - \mathbf{p}') + e^{-\frac{(p^2 - p'^2)^2}{\lambda^4}} \left[ \bar{V}_\lambda^{(1)}(\mathbf{p}, \mathbf{p}') + \bar{V}_\lambda^{(2)}(\mathbf{p}, \mathbf{p}') + \dots \right], \quad (35)$$

where

$$\bar{V}_\lambda^{(1)}(\mathbf{p}, \mathbf{p}') = -\frac{\alpha_{\lambda, i}}{(2\pi)^D}, \quad (36)$$

$$\bar{V}_\lambda^{(2)}(\mathbf{p}, \mathbf{p}') = \alpha_{\lambda, i}^2 F_s^{(2)}(\mathbf{p}, \mathbf{p}'), \quad (37)$$

$$\bar{V}_\lambda^{(n)}(\mathbf{p}, \mathbf{p}') = \alpha_{\lambda, i}^n F_s^{(n)}(\mathbf{p}, \mathbf{p}'). \quad (38)$$

Here  $\lambda$  is a momentum cutoff (as opposed to the energy cutoff discussed above) related to the flow parameter by  $s = 1/\lambda^4$ . The index  $i$  denotes the order of the calculation for the running coupling.

The renormalized hamiltonian can be used to compute eigenvalues and eigenstates. Since the hamiltonian is derived perturbatively we expect cutoff dependent errors in the observables. Formally, we can regroup the terms in the renormalized hamiltonian and write it as a momentum expansion, and the expansion parameters are analytic functions of the running coupling  $\alpha_\lambda$ . Expanding the operators  $F_s^{(n)}(\mathbf{p}, \mathbf{p}')$  in powers of  $p^2/\lambda^2$  we obtain

$$F_s^{(n)}(\mathbf{p}, \mathbf{p}') = z_0 + z_2 \frac{(p^2 + p'^2)}{2\lambda^2} + z_4 \frac{(p^4 + p'^4)}{4\lambda^4} + z'_4 \frac{p^2 p'^2}{2\lambda^4} + \dots, \quad (39)$$

where the  $z_i$ 's are constants. Regrouping the terms we obtain

$$H_\lambda(p, p') = p^2 \delta^{(D)}(\mathbf{p} - \mathbf{p}') + e^{-\frac{(p^2 - p'^2)^2}{\lambda^4}} \left[ g_0(\alpha_\lambda) + g_2(\alpha_\lambda) \frac{(p^2 + p'^2)}{2\lambda^2} + \dots \right], \quad (40)$$

where

$$g_i(\alpha_\lambda) = a_i \alpha_\lambda + b_i \alpha_\lambda^2 + \dots. \quad (41)$$

We can identify three interdependent sources of errors in the perturbative similarity renormalization group when the hamiltonian given by Eq. (40) is truncated and used to compute a physical quantity:

- a) errors introduced by the truncation of the hamiltonian at a given order in  $p^2/\lambda^2$ ;
- b) errors introduced by the truncation of the hamiltonian at a given order in the running coupling  $\alpha_{\lambda, i}$ , which correspond to the use of an approximation for the functions  $g_i$ ;
- c) errors introduced by the approximation for the running coupling  $\alpha_{\lambda, i}$ .

In the actual calculation using the hamiltonian given by Eq. (35) errors of type (a) do not appear directly because we do not truncate the operators that appear in the hamiltonian. However, errors of type (b) can be understood as coming from approximating the couplings in front of the operators in Eq. (40). Errors of type (c) appear in our calculations only because we do not fit the canonical coupling to data at each scale, but fix it at a given scale and evolve it perturbatively from that scale. The strategy we would use for a realistic theory (*e.g.*, QED and QCD) is the following:

- 1) Obtain the renormalized hamiltonian using the similarity transformation and coupling-coherence, truncating the hamiltonian at a given order in powers of  $\alpha_{\lambda,i}$ .
- 2) Fix the coupling  $\alpha_\lambda$  by fitting an observable (*e.g.*, a bound-state energy).
- 3) Evaluate other observables (*e.g.*, scattering phase shifts).

As pointed out before, the evaluation of scattering observables with the similarity hamiltonian with standard techniques is complicated and so in our examples we focus on the bound state errors. We fix the coupling at some scale using a given renormalization prescription and use the flow-equation to obtain the coupling as a function of the cutoff  $\lambda$  to a given order. We then perform a sequence of bound-state calculations with better approximations for the hamiltonian such that the errors in the bound-state energy are systematically reduced. Once the sources of errors are identified, it becomes relatively simple to analyze order-by-order how such errors scale with the cutoff  $\lambda$ . In principle, to completely eliminate the errors proportional to some power  $m$  in the momentum expansion we should use the similarity hamiltonian with the exact running coupling (renormalized to all orders) and include the contributions up to  $\mathcal{O}(p^m/\lambda^m)$  coming from all effective interactions (all orders in  $\alpha_\lambda$ ). Some details of this scaling analysis are presented later for the specific examples we work out. We emphasize again that in a realistic calculation we would fit the coupling  $\alpha_\lambda$  to an observable. This nonperturbative renormalization eliminates the dominant source of errors we display in SRG calculations in this paper. We choose to renormalize the coupling perturbatively in this paper because the only observable we compute is the single bound state energy of a delta-function potential, and fitting this energy would prevent us from displaying errors.

### 3 Two-Dimensional Delta-function Potential

We now consider the case of two nonrelativistic particles in two dimensions interacting via an attractive Dirac delta-function potential. The Schrödinger equation for relative motion in position space (with  $\hbar = 1$ ), can be written as:

$$-\nabla_{\mathbf{r}}^2 \Psi(\mathbf{r}) - \alpha_0 \delta^{(2)}(\mathbf{r}) \Psi(\mathbf{r}) = E \Psi(\mathbf{r}) . \quad (42)$$

Both the delta-function potential in two dimensions and the kinetic energy operator scale as  $1/r^2$ , therefore, the coupling  $\alpha_0$  is dimensionless. As a consequence, the hamiltonian is scale invariant (*i.e.*, there is no intrinsic energy scale) and we can anticipate the presence of logarithmic ultraviolet divergences, analogous to those appearing in QED and QCD. The problem requires renormalization. In this subsection we present the standard method that produces an exact solution analytically, using simple regularization and renormalization schemes [18].

#### 3.1 Exact Solution

We start with the Schrödinger equation in momentum space,

$$p^2 \Phi(\mathbf{p}) - \frac{\alpha_0}{(2\pi)^2} \int d^2 q \Phi(\mathbf{q}) = E \Phi(\mathbf{p}) , \quad (43)$$

where  $\Phi(\mathbf{p})$  is the Fourier transform of the position space wave-function,

$$\Phi(\mathbf{p}) = \frac{1}{2\pi} \int d^2 r \Psi(\mathbf{r}) e^{-i\mathbf{p}\cdot\mathbf{r}} . \quad (44)$$

As a consequence of scale invariance, if there is any negative energy solution to Eq. (43) then it will admit solutions for any  $E < 0$ . This corresponds to a continuum of bound states with energies

extending down to  $-\infty$ , so the system is not bounded from below. By rearranging the terms in the Schrödinger equation we obtain

$$\Phi(\mathbf{p}) = \frac{\alpha_0}{2\pi} \frac{\Psi(0)}{(p^2 + E_0)}, \quad (45)$$

where  $\Psi(0)$  is the position space wave-function at the origin and  $E_0 > 0$  is the binding energy.

To obtain the eigenvalue condition for the binding energy, we can integrate both sides of Eq. (45):

$$1 = \frac{\alpha_0}{2\pi} \int_0^\infty dp p \frac{1}{(p^2 + E_0)}. \quad (46)$$

The integral on the r.h.s. diverges logarithmically, so the problem is ill-defined.

The conventional way to deal with this problem is renormalization. First, we regulate the integral with a momentum cutoff, obtaining

$$1 = \frac{\alpha_0}{2\pi} \int_0^\Lambda dp p \frac{1}{(p^2 + E_0)} = \frac{\alpha_0}{4\pi} \ln \left( 1 + \frac{\Lambda^2}{E_0} \right), \quad (47)$$

so that

$$E_0 = \frac{\Lambda^2}{e^{-\frac{4\pi}{\alpha_0}} - 1}. \quad (48)$$

Clearly, if the coupling  $\alpha_0$  is fixed then  $E_0 \rightarrow \infty$  as  $\Lambda \rightarrow \infty$ . In order to eliminate the divergence and produce a finite, well-defined bound state we can renormalize the theory by demanding that the coupling runs with the cutoff  $\Lambda$  in such a way that the binding energy remains fixed as the cutoff is removed:

$$\alpha_0 \rightarrow \alpha_\Lambda = \frac{4\pi}{\ln \left( 1 + \frac{\Lambda^2}{E_0} \right)}. \quad (49)$$

The dimensionless renormalized running coupling  $\alpha_\Lambda$  that characterizes the strength of the interaction is therefore replaced by a new (dimensionful) parameter  $E_0 > 0$ , the binding energy of the system. This is a simple example of dimensional transmutation [38]: even though the original “bare” hamiltonian is scale invariant, the renormalization procedure leads to a scale that characterizes the physical observables. Note that  $E_0$  can be chosen arbitrarily, fixing the energy scale of the underlying (renormalized) theory. It is also interesting to note that the renormalized running coupling  $\alpha_\Lambda$  vanishes as  $\Lambda \rightarrow \infty$  and so the theory is asymptotically free.

This renormalized hamiltonian can be used to compute other observables. The usual prescription for the calculations is to obtain the solutions with the cutoff in place and then take the limit as the momentum cutoff is removed to  $\infty$ . If an exact calculation can be implemented, the final results should be independent of the regularization and renormalization schemes. As an example, we calculate the scattering wave function,

$$\Phi_k(\mathbf{p}) = \delta^{(2)}(\mathbf{p} - \mathbf{k}) + \frac{\alpha_\Lambda}{2\pi} \frac{\Psi(0)}{(p^2 - k^2 - i\epsilon)}, \quad (50)$$

where  $k = \sqrt{E}$ . Integrating both sides over  $\mathbf{p}$  with a cutoff  $\Lambda$  in place, we obtain

$$\Psi(0) = \frac{1}{2\pi} \left[ 1 - \frac{\alpha_\Lambda}{4\pi} \ln \left( 1 + \frac{\Lambda^2}{-k^2 - i\epsilon} \right) \right]^{-1}; \quad (51)$$

thus,

$$\alpha_\Lambda \Psi(0) = \frac{1}{2\pi} \left[ \frac{1}{4\pi} \ln \left( 1 + \frac{\Lambda^2}{E_0} \right) - \frac{1}{4\pi} \ln \left( 1 + \frac{\Lambda^2}{-k^2 - i\epsilon} \right) \right]^{-1}. \quad (52)$$

In the limit  $\Lambda \rightarrow \infty$  we obtain:

$$\alpha_\Lambda \Psi(0) = \frac{2}{\ln \left( \frac{k^2}{E_0} \right) - i\pi}. \quad (53)$$

The resulting scattering wave function is then given by

$$\Phi_k(\mathbf{p}) = \delta^{(2)}(\mathbf{p} - \mathbf{k}) + \frac{1}{2\pi} \frac{2}{(p^2 - k^2 - i\epsilon)} \left[ \ln \left( \frac{k^2}{E_0} \right) - i\pi \right]^{-1}. \quad (54)$$

It is important to note that only S-wave scattering occurs, corresponding to zero angular momentum states. For the higher waves the centrifugal barrier completely screens the delta-function potential and the non-zero angular momentum scattering states are free states.

The same prescription can be used to evaluate the T-matrix or the K-matrix. For the T-matrix, the Lippmann-Schwinger equation with the renormalized potential is given by:

$$T(\mathbf{p}, \mathbf{p}'; k) = V(\mathbf{p}, \mathbf{p}') + \int d^2q \frac{V(\mathbf{p}, \mathbf{q})}{k^2 - q^2 + i\epsilon} T(\mathbf{q}, \mathbf{p}'; k). \quad (55)$$

Since only S-wave scattering takes place we can integrate out the angular variable, obtaining

$$T^{(l=0)}(p, p'; k) = V^{(l=0)}(p, p') + \int_0^\Lambda dq q \frac{V^{(l=0)}(p, q)}{k^2 - q^2 + i\epsilon} T^{(l=0)}(q, p'; k), \quad (56)$$

where

$$V^{(l=0)}(p, p') = -\frac{\alpha\Lambda}{2\pi}. \quad (57)$$

The Lippmann-Schwinger equation for the “on-shell” T-matrix is given by:

$$T^{(l=0)}(k) = -\frac{\alpha\Lambda}{2\pi} - \frac{\alpha\Lambda}{2\pi} T^{(l=0)}(k) \int_0^\Lambda dq q \frac{1}{k^2 - q^2 + i\epsilon}. \quad (58)$$

Solving this equation and taking the limit  $\Lambda \rightarrow \infty$ , we obtain the exact “on-shell” T-matrix:

$$T_0(k) = -\frac{2}{\ln \left( \frac{k^2}{E_0} \right) - i\pi}. \quad (59)$$

Here and in what follows we drop the superscript and use the subscript 0 to denote the exact quantities.

In the same way, the S-wave Lippmann-Schwinger equation for the K-matrix is given by

$$K(p, p'; k) = V(p, p') + \mathcal{P} \int_0^\Lambda dq q \frac{V(p, q)}{k^2 - q^2} K(q, p'; k), \quad (60)$$

and the exact “on-shell” K-matrix is given by

$$K_0(k) = -\frac{2}{\ln \left( \frac{k^2}{E_0} \right)}. \quad (61)$$

The “on-shell” K-matrix and T-matrix are related by

$$K_0(k) = \frac{T_0(k)}{1 - \frac{i\pi}{2} T_0(k)}. \quad (62)$$

Using either

$$k \cot \delta_0(k) - ik = -\frac{2k}{\pi} \frac{1}{T_0(k)}, \quad (63)$$

or

$$k \cot \delta_0(k) = -\frac{2k}{\pi} \frac{1}{K_0(k)}, \quad (64)$$

we can obtain the exact phase-shifts:

$$\cot \delta_0 = \frac{1}{\pi} \ln \left( \frac{k^2}{E_0} \right). \quad (65)$$

### 3.2 Similarity Renormalization Group Approach

In the two-dimensional case the canonical hamiltonian in momentum space with a delta-function potential can be written as

$$H(\mathbf{p}, \mathbf{p}') = h(\mathbf{p}, \mathbf{p}') + V(\mathbf{p}, \mathbf{p}') , \quad (66)$$

where  $h(\mathbf{p}, \mathbf{p}') = p^2 \delta^{(2)}(\mathbf{p} - \mathbf{p}')$  corresponds to the free hamiltonian and  $V(\mathbf{p}, \mathbf{p}') = -\alpha_0/(2\pi)^2$  corresponds to the Fourier transform of the delta-function potential.

Integrating out the angular variable, the flow equation obtained with Wegner's transformation in terms of matrix elements in the basis of free states is given by

$$\frac{dV_s(p, p')}{ds} = -(p^2 - p'^2)^2 V_s(p, p') - \int_0^\infty dk k (2k^2 - p^2 - p'^2) V_s(p, k) V_s(k, p') . \quad (67)$$

In principle, we can set the boundary condition at  $s = 0$  (no cutoff), i.e.,

$$H_{s=0}(p, p') = H(p, p') = p^2 \delta^{(1)}(p - p') - \frac{\alpha_0}{2\pi} . \quad (68)$$

However, the hamiltonian with no cutoff produces logarithmic divergences, requiring renormalization. As we will see, the boundary condition must be imposed at some other point, leading to dimensional transmutation [38]. The reduced interaction  $\bar{V}_s(p, p')$  is defined such that

$$V_s(p, p') = e^{-s(p^2 - p'^2)^2} \bar{V}_s(p, p') . \quad (69)$$

Assuming that  $h$  is cutoff independent we obtain the flow equation for the reduced interaction,

$$\begin{aligned} \frac{d\bar{V}_s}{ds} = & -e^{-2s p^2 p'^2} \int_0^\infty dk k (2k^2 - p^2 - p'^2) e^{-2s[k^4 - k^2(p^2 + p'^2)]} \\ & \times \bar{V}_s(p, k) \bar{V}_s(k, p') . \end{aligned} \quad (70)$$

This equation is solved using a perturbative expansion, starting with

$$\bar{V}_s^{(1)}(p, p') = -\frac{\alpha_s}{2\pi} . \quad (71)$$

We assume a coupling-coherent solution in the form of an expansion in powers of  $\alpha_s/2\pi$ , satisfying the constraint that the operators  $F_s^{(n)}(p, p')$  vanish when  $p = p' = 0$ ,

$$\bar{V}_s(p, p') = -\frac{\alpha_s}{2\pi} + \sum_{n=2}^{\infty} \left(\frac{\alpha_s}{2\pi}\right)^n F_s^{(n)}(p, p') . \quad (72)$$

Note that the expansion parameter is  $\alpha_s/2\pi$ .

Using the solution Eq. (72) in Eq. (70) we obtain

$$\begin{aligned} \frac{d\bar{V}_s}{ds} = & -\frac{1}{(2\pi)^2} \frac{d\alpha_s}{ds} + \sum_{n=2}^{\infty} \frac{1}{(2\pi)^n} \left[ n \alpha_s^{n-1} \frac{d\alpha_s}{ds} F_s^{(n)}(p, p') + \alpha_s^n \frac{dF_s^{(n)}(p, p')}{ds} \right] \\ = & \int_0^\infty dk k (2k^2 - p^2 - p'^2) e^{-2s[p^2 p'^2 + k^4 - k^2(p^2 + p'^2)]} \\ \times & \left[ -\frac{\alpha_s}{2\pi} + \sum_{n=2}^{\infty} \left(\frac{\alpha_s}{2\pi}\right)^n F_s^{(n)}(p, k) \right] \left[ -\frac{\alpha_s}{2\pi} + \sum_{m=2}^{\infty} \left(\frac{\alpha_s}{2\pi}\right)^m F_s^{(m)}(k, p') \right] . \end{aligned} \quad (73)$$

This equation is solved iteratively order-by-order in  $\alpha_s/2\pi$ . Again, if  $\alpha_s/2\pi$  is small the operator  $\bar{V}_s^{(1)}(p, p')$  can be identified as the dominant term in the expansion of  $\bar{V}_s(p, p')$  in powers of  $p$  and  $p'$ . In the  $D = 2$  case this operator corresponds to a marginal operator (since the coupling is dimensionless and there is no implicit mass scale). The higher-order terms correspond to irrelevant operators.



At second-order we have

$$-\frac{1}{2\pi} \frac{d\alpha_s}{ds} + \frac{1}{(2\pi)^2} \alpha_s^2 \frac{dF_s^{(2)}(p, p')}{ds} = -\alpha_s^2 I_s^{(2)}(p, p'), \quad (74)$$

where

$$\begin{aligned} I_s^{(2)}(p, p') &= \frac{1}{(2\pi)^2} \int_0^\infty dk k (2k^2 - p^2 - p'^2) e^{-2s[p^2 p'^2 + k^4 - k^2(p^2 + p'^2)]} \\ &= \frac{1}{(2\pi)^2} \frac{e^{-2s p^2 p'^2}}{4s}. \end{aligned} \quad (75)$$

The equation for  $\alpha_s$  is obtained by taking the limit  $(p, p') \rightarrow 0$ ,

$$\frac{1}{2\pi} \frac{d\alpha_s}{ds} = \alpha_s^2 I_s^{(2)}(0, 0), \quad (76)$$

where

$$I_s^{(2)}(0, 0) = \frac{1}{(2\pi)^2} \frac{1}{4s}. \quad (77)$$

Integrating Eq. (76) from  $s_0$  to  $s$ ,

$$\alpha_{s,2} = \frac{\alpha_{s_0}}{1 - \frac{\alpha_{s_0}}{8\pi} \ln\left(\frac{s}{s_0}\right)}. \quad (78)$$

In terms of the cutoff  $\lambda$  we obtain

$$\alpha_{\lambda,2} = \frac{\alpha_{\lambda_0}}{1 + \frac{\alpha_{\lambda_0}}{2\pi} \ln\left(\frac{\lambda}{\lambda_0}\right)}. \quad (79)$$

In principle, knowing the value of  $\alpha_{s_0}$  for a given  $s_0$  we can determine the running coupling  $\alpha_s$  for any  $s$ . Since we cannot choose  $s_0 = 0$  ( $\lambda_0 = \infty$ ), to use Eq. (78) we must specify a renormalization prescription that allows us to fix the coupling at some finite non-zero value of  $s_0$ . We discuss this issue in detail later in this subsection.

The equation for  $F_s^{(2)}(p, p')$  is given by

$$\frac{1}{(2\pi)^2} \frac{dF_s^{(2)}(p, p')}{ds} = I_s^{(2)}(0, 0) - I_s^{(2)}(p, p'). \quad (80)$$

Integrating from  $s_0$  to  $s$  we obtain

$$\begin{aligned} F_s^{(2)}(p, p') &= \int_{s_0}^s ds' \frac{[1 - e^{-2s' p^2 p'^2}]}{4s'} \\ &= \frac{1}{4} \left[ \ln\left(\frac{s}{s_0}\right) - \text{Ei}(-2s p^2 p'^2) + \text{Ei}(-2s_0 p^2 p'^2) \right]. \end{aligned} \quad (81)$$

Insisting that  $F_s^{(2)}(p, p') = 0$  when  $p, p' = 0$  we obtain

$$F_s^{(2)}(p, p') = \frac{1}{4} [\gamma + \ln(2s p^2 p'^2) - \text{Ei}(-2s p^2 p'^2)]. \quad (82)$$

At third-order we have

$$\begin{aligned} -\frac{1}{2\pi} \frac{d\alpha_s}{ds} + \frac{1}{(2\pi)^2} \alpha_s^2 \frac{dF_s^{(2)}(p, p')}{ds} &+ \frac{2}{(2\pi)^2} \alpha_s \frac{d\alpha_s}{ds} F_s^{(2)}(p, p') + \frac{1}{(2\pi)^3} \alpha_s^3 \frac{dF_s^{(3)}(p, p')}{ds} \\ &= -\alpha_s^2 I_s^{(2)}(p, p') + \alpha_s^3 I_s^{(3)}(p, p'), \end{aligned} \quad (83)$$

where

$$I_s^{(3)}(p, p') = \frac{1}{(2\pi)^3} \int_0^\infty dk k (2k^2 - p^2 - p'^2) e^{-2s[p^2 p'^2 + k^4 - k^2(p^2 + p'^2)]} \\ \times \left[ F_s^{(2)}(p, k) + F_s^{(2)}(k, p') \right]. \quad (84)$$

In the limit  $p, p' \rightarrow 0$  we obtain:

$$\frac{1}{2\pi} \frac{d\alpha_s}{ds} = \alpha_s^2 I_s^{(2)}(0, 0) - \alpha_s^3 I_s^{(3)}(0, 0), \quad (85)$$

where

$$I_s^{(3)}(0, 0) = \frac{1}{(2\pi)^3} \int_0^\infty dk 2k^3 e^{-2sk^4} \left[ F_s^{(2)}(0, k) + F_s^{(2)}(k, 0) \right]. \quad (86)$$

Since  $F_s^{(2)}(k, 0) = F_s^{(2)}(0, k) = 0$ , the term proportional to  $\alpha_s^3$  in Eq. (85) is zero.

The equation for  $F_s^{(3)}(p, p')$  is given by

$$\frac{1}{(2\pi)^3} \frac{dF_s^{(3)}(p, p')}{ds} = -\frac{1}{\pi} I_s^{(2)}(0, 0) F_s^{(2)}(p, p') + I_s^{(3)}(p, p'). \quad (87)$$

To obtain  $F_s^{(3)}(p, p')$  the integrals in  $k$  and  $s$  must be evaluated numerically.

At fourth-order we obtain

$$-\frac{1}{2\pi} \frac{d\alpha_s}{ds} + \frac{1}{(2\pi)^2} \alpha_s^2 \frac{dF_s^{(2)}(p, p')}{ds} + \frac{2}{(2\pi)^2} \alpha_s \frac{d\alpha_s}{ds} F_s^{(2)}(p, p') \\ + \frac{1}{(2\pi)^3} \alpha_s^3 \frac{dF_s^{(3)}(p, p')}{ds} + \frac{3}{(2\pi)^3} \alpha_s^2 \frac{d\alpha_s}{ds} F_s^{(3)}(p, p') + \frac{1}{(2\pi)^4} \alpha_s^4 \frac{dF_s^{(4)}(p, p')}{ds} \\ = -\alpha_s^2 I_s^{(2)}(p, p') + \alpha_s^3 I_s^{(3)}(p, p') + \alpha_s^4 I_s^{(4)}(p, p'), \quad (88)$$

where

$$I_s^{(4)}(p, p') = \frac{1}{(2\pi)^4} \int_0^\infty dk k (2k^2 - p^2 - p'^2) e^{-2s[p^2 p'^2 + k^4 - k^2(p^2 + p'^2)]} \\ \times \left[ F_s^{(3)}(p, k) + F_s^{(3)}(k, p') + F_s^{(2)}(p, k) F_s^{(2)}(k, p') \right]. \quad (89)$$

In the limit  $p, p' \rightarrow 0$  we obtain:

$$\frac{1}{2\pi} \frac{d\alpha_s}{ds} = \alpha_s^2 I_s^{(2)}(0, 0) - \alpha_s^4 I_s^{(4)}(0, 0), \quad (90)$$

where

$$I_s^{(4)}(0, 0) = \frac{1}{2\pi} \int_0^\infty dk k 2k^2 e^{-2sk^4} \left[ F_s^{(3)}(0, k) + F_s^{(3)}(k, 0) \right]. \quad (91)$$

For dimensional reasons Eq. (90) takes the form

$$\frac{d\alpha_s}{ds} = \frac{B_2}{s} \alpha_s^2 - \frac{B_4}{s} \alpha_s^4, \quad (92)$$

where  $B_2 = \frac{1}{8\pi}$  and  $B_4$  can be obtained by evaluating  $I_s^{(4)}(0, 0)$  (numerically) for  $s = 1$ . In terms of the cutoff  $\lambda$  we obtain:

$$\frac{d\alpha_\lambda}{d\lambda} = \frac{1}{\lambda^4} B_2 \alpha_\lambda^2 - \frac{1}{\lambda^4} B_4 \alpha_\lambda^4. \quad (93)$$

Integration of Eq. (90) leads to a transcendental equation which is solved numerically in order to obtain the running coupling  $\alpha_{s,4}$ .

Qualitatively, the errors are expected to be a combination of inverse powers of  $\lambda$  and powers (or inverse powers) of  $\ln(\lambda)$  coming from the perturbative expansion in powers of  $\alpha_\lambda$  for the coefficients of the irrelevant operators and the perturbative approximation for  $\alpha_\lambda$ .

As pointed out above, to completely determine the renormalized hamiltonian at a given order we need to specify the coupling at some scale  $\lambda_0$ . The simplest way is to choose a value for the ‘exact’  $\alpha_{\lambda_0}$ . Formally, this fixes the underlying theory; *i.e.*, if we had the exact hamiltonian (to all orders) we could obtain the exact values for all observables. However, since the hamiltonian is derived perturbatively and must be truncated at some order in practical calculations, we can only obtain approximate cutoff-dependent results for the observables. Moreover, in this case the errors cannot be directly evaluated, since the exact values remain unknown. As an example, we calculate the bound-state energy choosing  $\alpha_{\lambda_0} = 1.45$  at  $\lambda_0 = 100$ . In Fig. 2 we show the binding-energy as a function of the cutoff  $\lambda$  using the following approximations for the interaction:

(a) marginal operator with coupling ( $\alpha_{\lambda_0}$ ),

$$V_\lambda(p, p') = -\frac{\alpha_{\lambda_0}}{2\pi} e^{-\frac{(p^2 - p'^2)^2}{\lambda^4}}; \quad (94)$$

(b) marginal operator with running coupling renormalized to second-order ( $\alpha_{\lambda,2}$ ),

$$V_\lambda(p, p') = -\frac{\alpha_{\lambda,2}}{2\pi} e^{-\frac{(p^2 - p'^2)^2}{\lambda^4}}; \quad (95)$$

(c) marginal operator plus second-order irrelevant operator with running coupling renormalized to second-order ( $\alpha_{\lambda,2}, F_\lambda^{(2)}$ ),

$$V_\lambda(p, p') = \left[ -\frac{\alpha_{\lambda,2}}{2\pi} + \left( \frac{\alpha_{\lambda,2}}{2\pi} \right)^2 F_\lambda^{(2)}(p, p') \right] e^{-\frac{(p^2 - p'^2)^2}{\lambda^4}}; \quad (96)$$

(d) marginal operator with running coupling renormalized to fourth-order ( $\alpha_{\lambda,4}$ ),

$$V_\lambda(p, p') = -\frac{\alpha_{\lambda,4}}{2\pi} e^{-\frac{(p^2 - p'^2)^2}{\lambda^4}}; \quad (97)$$

(e) marginal operator plus second-order irrelevant operator with running coupling renormalized to fourth-order ( $\alpha_{\lambda,4}, F_\lambda^{(2)}$ ),

$$V_\lambda(p, p') = \left[ -\frac{\alpha_{\lambda,4}}{2\pi} + \left( \frac{\alpha_{\lambda,4}}{2\pi} \right)^2 F_\lambda^{(2)}(p, p') \right] e^{-\frac{(p^2 - p'^2)^2}{\lambda^4}}. \quad (98)$$

We see that as the approximation is improved the cutoff dependence is reduced. As  $\lambda \rightarrow \infty$  all curves should approach the same binding-energy, which corresponds to the exact value, and as  $\lambda$  becomes small the perturbative approximation breaks down.

A similar prescription is to find  $\alpha_{\lambda_0}$  at  $\lambda_0$  that produces a given binding-energy,  $E_0$ . Since the fitting is implemented using a truncated hamiltonian, this  $\alpha_{\lambda_0}$  is an approximation that becomes more accurate if we use a larger  $\lambda_0$  and/or include higher order operators. Although in this case we can evaluate the errors, the scaling analysis becomes complicated as  $\lambda \rightarrow \lambda_0$  because at this point we force the energy to be the exact value and so the error is zero.

As an example, we calculate the bound-state energy when the coupling is fixed at  $\lambda_0 = 100$  to give  $E_0 = 1$ . In Fig. 3 we show the ‘errors’ in the binding energy using the same approximations listed above for the potential. As expected, all error lines drop abruptly to zero when  $\lambda \rightarrow \lambda_0$ , where the running coupling is chosen to fit what we define to be the exact binding energy. Away from this point we can analyze the errors. With the hamiltonian (a) (unrenormalized) we obtain a dominant

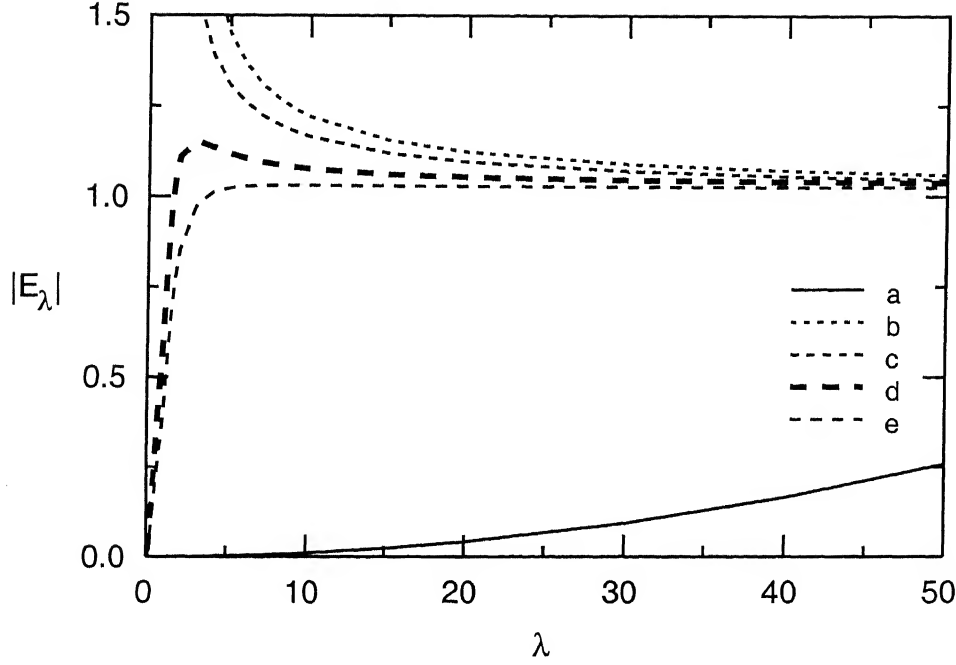


Figure 2: The binding energy for the two-dimensional delta-function potential with various approximations for the SRG hamiltonian. The exact theory is fixed by choosing  $\alpha_0 = 1.45$  at  $\lambda_0 = 100$ .

error that scales like  $\ln(\lambda_0/\lambda)$ , corresponding to the leading order error. With the renormalized hamiltonian (b) the dominant errors scale like  $[\ln(\lambda_0/\lambda)]^{-2}$ , indicating the elimination of the leading order logarithmic errors. With the hamiltonian (c) there is a small shift, but no significant change in the error scaling. The added irrelevant operator may remove errors of order  $(\lambda_0/\lambda)^2$ , but these are smaller than the remaining  $[\ln(\lambda_0/\lambda)]^{-2}$  errors. With hamiltonians (d) and (e) in the range of intermediate cutoffs ( $E_0 \ll \lambda^2 \ll \lambda_0^2$ ) there is also only a shift in the errors. The dips in (d) and (e) correspond to values of  $\lambda$  where the binding energy equals the exact value.

This behavior is a perturbative artifact that can be understood in the following way. Consider the Schrödinger equation with potential (d). Rescaling the momenta  $p \rightarrow \lambda \tilde{p}$  we obtain

$$\tilde{p}^2 \tilde{\Phi}(\tilde{p}) - \frac{\alpha_{\lambda,4}}{2\pi} \int_0^\infty d\tilde{q} \tilde{q} e^{(\tilde{p}^2 - \tilde{q}^2)} \tilde{\Phi}(\tilde{q}) = \frac{E_\lambda}{\lambda^2} \tilde{\Phi}(\tilde{p}). \quad (99)$$

and

$$E_\lambda = \lambda^2 \frac{\left[ \int_0^\infty d\tilde{p} \tilde{p} \left( \tilde{p}^2 |\tilde{\Phi}(\tilde{p})|^2 \right) - \frac{\alpha_{\lambda,4}}{2\pi} \int_0^\infty d\tilde{p} \tilde{p} \int_0^\infty d\tilde{q} \tilde{q} e^{(\tilde{p}^2 - \tilde{q}^2)} \tilde{\Phi}(\tilde{p}) \tilde{\Phi}(\tilde{q}) \right]}{\int_0^\infty d\tilde{p} \tilde{p} |\tilde{\Phi}(\tilde{p})|^2}. \quad (100)$$

As shown in Fig. 4, the coupling renormalized to fourth-order,  $\alpha_{\lambda,4}$ , approximately freezes for small  $\lambda$  and as a consequence the bound-state energy scales like  $E_\lambda \simeq \lambda^2 \times \text{constant}$  eventually becoming equal to the exact value and then deviating again. With the hamiltonian (e) the behavior is similar, with the dip occurring at a different value of  $\lambda$  because of the irrelevant operator. For small values of  $\lambda$  the lines converge, indicating the breakdown of the perturbative expansion.

An alternative prescription is to use the potential derived in subsection 5.1 as the starting point for the similarity transformation. We introduce a large momentum cutoff  $\Lambda$ , define

$$\alpha_{s_0=0} = \alpha_\Lambda = \frac{4\pi}{\ln\left(1 + \frac{\Lambda^2}{E_0}\right)}, \quad (101)$$

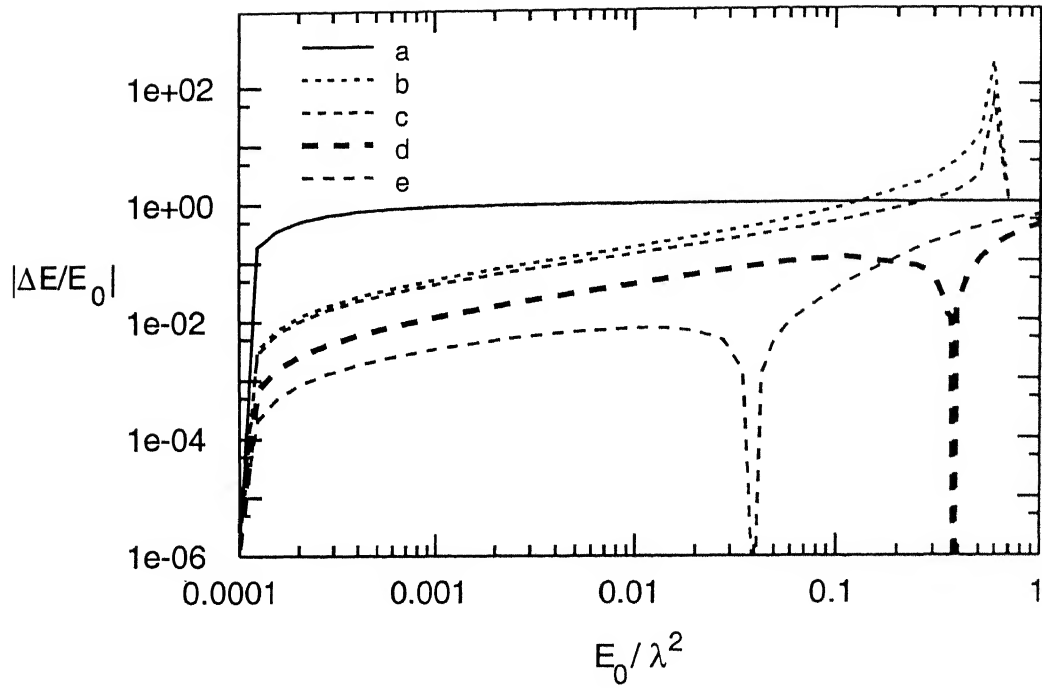


Figure 3: The SRG errors in the binding energy for the two-dimensional delta-function potential using various approximations for the SRG hamiltonian. The exact theory is fixed by choosing  $E_0 = 1$  at  $\lambda_0 = 100$ .

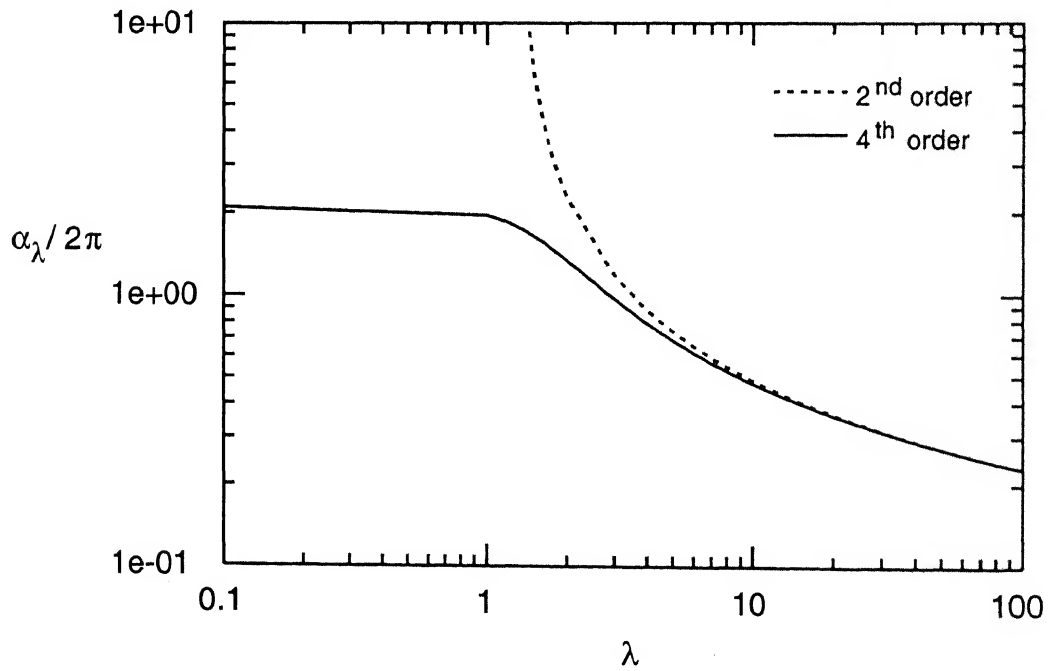


Figure 4: The SRG running coupling for the two-dimensional delta-function potential obtained with  $\alpha_{\lambda_0}$  at  $\lambda_0 = 100$  fixed to fit  $E_0 = 1$ .

and set all irrelevant operators to zero at  $s_0 = 0$ . Note that the coupling  $\alpha_{\lambda_0}$  is fixed at  $\lambda_0 = \infty$  by fitting the exact binding energy. With this definition the similarity hamiltonian with no similarity cutoff becomes well-defined and we can set all of the similarity transformation boundary conditions at  $s = 0$ . The previous derivation remains essentially the same. The only modification is that all integrals over momentum are cut off ( $p \leq \Lambda$ ).

At second-order we have

$$-\frac{1}{2\pi} \frac{d\alpha_{s,\Lambda}}{ds} + \frac{1}{(2\pi)^2} \alpha_{s,\Lambda}^2 \frac{dF_{s,\Lambda}^{(2)}(p, p')}{ds} = -\alpha_{s,\Lambda}^2 I_{s,\Lambda}^{(2)}(p, p') \quad (102)$$

where

$$\begin{aligned} I_{s,\Lambda}^{(2)}(p, p') &= \frac{1}{(2\pi)^2} \int_0^\Lambda dk k (2k^2 - p^2 - p'^2) e^{-2s[p^2 p'^2 + k^4 - k^2(p^2 + p'^2)]} \\ &= \frac{1}{(2\pi)^2} \frac{e^{-2s p^2 p'^2}}{4s}. \end{aligned} \quad (103)$$

The resulting second-order running coupling and irrelevant operator are given respectively by

$$\alpha_{s,\Lambda,2} = \frac{\alpha_\Lambda}{1 - \frac{\alpha_\Lambda}{8\pi} [\gamma + \ln(2s\Lambda^4) - \text{Ei}(-2s\Lambda^4)]} \quad (104)$$

and

$$\begin{aligned} F_{s,\Lambda}^{(2)}(p, p') &= \frac{1}{4} [\gamma + \ln(2s p^2 p'^2) - \text{Ei}(-2s p^2 p'^2)] \\ &+ \frac{1}{4} [\gamma + \ln(2s\Lambda^4) - \text{Ei}(-2s\Lambda^4)] \\ &- \frac{1}{4} [\gamma + \ln(s [(p^2 - \Lambda^2)^2 + (p'^2 - \Lambda^2)^2 - (p^2 - p'^2)^2]) \\ &\quad - \text{Ei}(-s [(p^2 - \Lambda^2)^2 + (p'^2 - \Lambda^2)^2 - (p^2 - p'^2)^2])] . \end{aligned} \quad (105)$$

At third-order we have

$$\begin{aligned} -\frac{1}{2\pi} \frac{d\alpha_{s,\Lambda}}{ds} + \frac{1}{(2\pi)^2} \alpha_{s,\Lambda}^2 \frac{dF_{s,\Lambda}^{(2)}(p, p')}{ds} + \frac{2}{(2\pi)^2} \alpha_{s,\Lambda} \frac{d\alpha_{s,\Lambda}}{ds} F_{s,\Lambda}^{(2)}(p, p') \\ + \frac{1}{(2\pi)^3} \alpha_{s,\Lambda}^3 \frac{dF_{s,\Lambda}^{(3)}(p, p')}{ds} = -\alpha_{s,\Lambda}^2 I_{s,\Lambda}^{(2)}(p, p') + \alpha_{s,\Lambda}^3 I_{s,\Lambda}^{(3)}(p, p') , \end{aligned} \quad (106)$$

where

$$\begin{aligned} I_{s,\Lambda}^{(3)}(p, p') &= \frac{1}{(2\pi)^3} \int_0^\Lambda dk k (2k^2 - p^2 - p'^2) e^{-2s[p^2 p'^2 + k^4 - k^2(p^2 + p'^2)]} \\ &\times [F_s^{(2)}(p, k) + F_s^{(2)}(k, p')] . \end{aligned} \quad (107)$$

In this case,

$$\begin{aligned} F_{s,\Lambda}^{(2)}(0, k) &= \frac{1}{4} [\gamma + \ln(2s\Lambda^4) - \text{Ei}(-2s\Lambda^4)] \\ &= \frac{1}{4} [\gamma + \ln(2s\Lambda^4 - 2sk^2\Lambda^2) - \text{Ei}(-2s\Lambda^4 + 2sk^2\Lambda^2)] , \end{aligned} \quad (108)$$

and so

$$\begin{aligned} I_{s,\Lambda}^{(3)}(0, 0) &= \frac{1}{(2\pi)^3} \frac{1}{4} \int_0^\Lambda dk 4k^3 e^{-2sk^4} ([\gamma + \ln(2s\Lambda^4) - \text{Ei}(-2s\Lambda^4)] \\ &- [\gamma + \ln(2s\Lambda^4 - 2sk^2\Lambda^2) - \text{Ei}(-2s\Lambda^4 + 2sk^2\Lambda^2)]) . \end{aligned} \quad (109)$$

Since  $I_{s,\Lambda}^{(3)}(0,0) \neq 0$  the term proportional to  $\alpha_{s,\Lambda}^3$  in Eq. (85) does not vanish. To obtain  $\alpha_{s,\Lambda,3}$  we evaluate  $I_{s,\Lambda}^{(3)}(0,0)$  and solve Eq. (85) numerically.

The equation for  $F_{s,\Lambda}^{(3)}(p,p')$  is given by

$$\frac{1}{(2\pi)^3} \frac{dF_{s,\Lambda}^{(3)}(p,p')}{ds} = -\frac{1}{\pi} I_{s,\Lambda}^{(2)}(0,0) F_{s,\Lambda}^{(2)}(p,p') + I_{s,\Lambda}^{(3)}(p,p'). \quad (110)$$

To obtain  $F_{s,\Lambda}^{(3)}(p,p')$  the integrals over  $k$  and  $s$  must be evaluated numerically. In the limit  $s\Lambda^4 \rightarrow \infty$  with  $s$  fixed at some non-zero value

$$F_{s,\Lambda}^{(2)}(p,p') \rightarrow \frac{1}{4} [\gamma + \ln(2s p^2 p'^2) - \text{Ei}(-2s p^2 p'^2)] , \quad (111)$$

$$I_{s,\Lambda}^{(3)}(0,0) \rightarrow 0 , \quad (112)$$

and  $\alpha_{s,\Lambda}$  becomes indeterminate, requiring the coupling to be fixed at some  $s_0 \neq 0$ . In this way, we recover the result of the previous prescription.

Although less trivial, this prescription allows a more transparent error analysis. We can also extend the calculation to larger values of the similarity cutoff,  $\lambda$ , and analyze the errors in a different scaling regime. In Fig. 5 we show the errors in the binding-energy obtained using the following approximations for the potential with  $\Lambda = 50$ :

(a) marginal operator with coupling  $(\alpha_{\lambda_0})$ ,

$$V_{\lambda,\Lambda}(p,p') = -\frac{\alpha_\Lambda}{2\pi} e^{-\frac{(p^2-p'^2)^2}{\lambda^4}} ; \quad (113)$$

(b) marginal operator with running coupling renormalized to second-order  $(\alpha_{\lambda,\Lambda,2})$ ,

$$V_{\lambda,\Lambda}(p,p') = -\frac{\alpha_{\lambda,\Lambda,2}}{2\pi} e^{-\frac{(p^2-p'^2)^2}{\lambda^4}} ; \quad (114)$$

(c) marginal operator plus second-order irrelevant operator with running coupling renormalized to second-order  $(\alpha_{\lambda,\Lambda,2}, F_\lambda^{(2)})$ ,

$$V_{\lambda,\Lambda}(p,p') = \left[ -\frac{\alpha_{\lambda,\Lambda,2}}{2\pi} + \left( \frac{\alpha_{\lambda,\Lambda,2}}{2\pi} \right)^2 F_{\lambda,\Lambda}^{(2)}(p,p') \right] e^{-\frac{(p^2-p'^2)^2}{\lambda^4}} ; \quad (115)$$

(d) marginal operator with running coupling renormalized to third-order  $(\alpha_{\lambda,\Lambda,3})$ ,

$$V_{\lambda,\Lambda}(p,p') = -\frac{\alpha_{\lambda,\Lambda,3}}{2\pi} e^{-\frac{(p^2-p'^2)^2}{\lambda^4}} ; \quad (116)$$

(e) marginal operator plus second-order irrelevant operator with running coupling renormalized to third-order  $\alpha_{\lambda,\Lambda,3}, F_\lambda^{(2)}$ ,

$$V_{\lambda,\Lambda}(p,p') = \left[ -\frac{\alpha_{\lambda,\Lambda,3}}{2\pi} + \left( \frac{\alpha_{\lambda,\Lambda,3}}{2\pi} \right)^2 F_{\lambda,\Lambda}^{(2)}(p,p') \right] e^{-\frac{(p^2-p'^2)^2}{\lambda^4}} . \quad (117)$$

There are clearly two distinct scaling regions when an additional large momentum cutoff  $\Lambda$  is placed on the initial matrix and a similarity cutoff is then applied. When the similarity cutoff is larger than  $\Lambda$ , we see power-law improvement resulting from the addition of irrelevant operators. Curves (a), (b), and (d) all have the same slope. None of these hamiltonians contains irrelevant operators, but the marginal coupling differs in each. All results become exact as the similarity cutoff goes to infinity, and these curves are close to one another because the coupling runs little

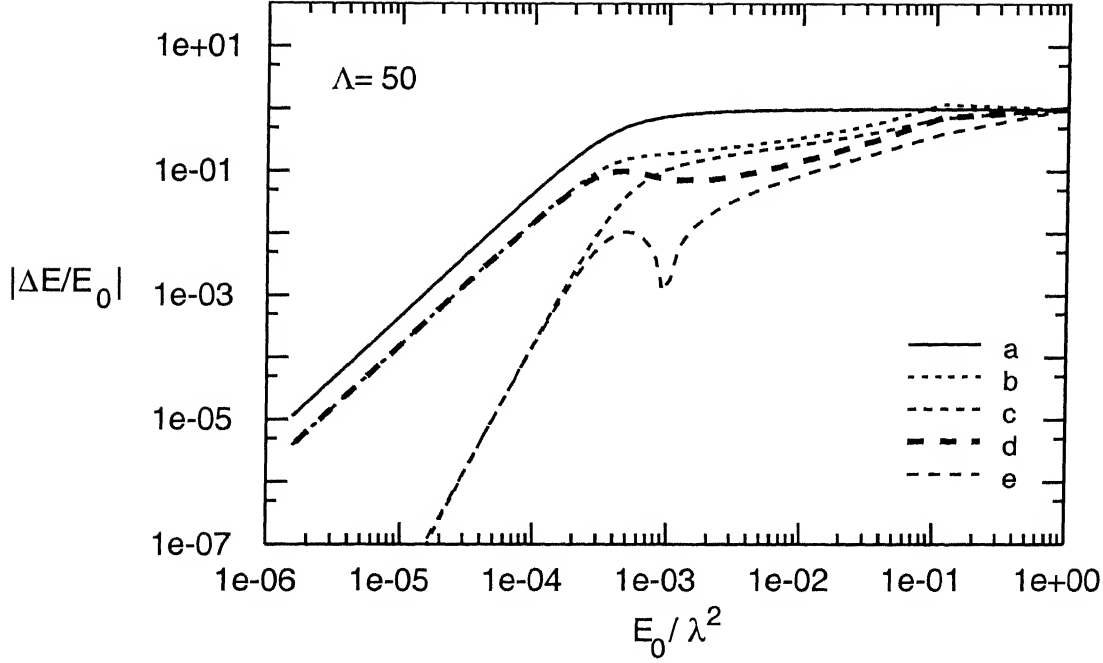


Figure 5: The SRG errors in the binding energy for the two-dimensional delta-function potential using various approximations for the similarity hamiltonian. The exact theory is fixed by regulating the “bare hamiltonian” using a sharp momentum cutoff,  $\Lambda$ , and letting the bare coupling depend on  $\Lambda$  such that the binding energy is fixed. We use  $\Lambda = 50$  and  $E_0 = 1$ .

in this region. Curves (c) and (e) show that there is a power-law improvement when irrelevant operators are added, and that once again when the similarity cutoff is larger than  $\Lambda$  an improvement in the running coupling makes little difference. Even though the coupling in front of this operator is approximated by the first term in an expansion in powers of the running coupling, the coupling is sufficiently small that this approximation works well and the operator eliminates most of the leading power-law error in curves (a), (b), and (d).

When the similarity cutoff become smaller than  $\Lambda$  we see a crossover to a more complicated scaling regime that resembles the SRG scaling discussed above. The error displayed by curve (a) approaches 100%, while the running coupling introduced in curve (b) reduces the error to an inverse logarithm. Improving the running coupling in curve (d) further reduces the error, and we see that curve (c) crosses curve (d) at a point where improving the running coupling becomes more important than adding irrelevant operators. As above, the best results require us to both improve the running coupling by adding third-order corrections and add the second-order irrelevant operators. In no case do we achieve power-law improvement, because as we have discussed there are always residual inverse logarithmic errors. Had we fit the running coupling to data, as we would do in a realistic calculation, we would obtain power-law improvement and the residual error would be proportional to an inverse power of the cutoff times an inverse power of the logarithm of the cutoff.

In Fig. 6 we show the running coupling at 2nd and 3rd order. Although the 3rd order corrections are small for all  $\lambda$  and vanish when  $\Lambda \rightarrow \infty$ , the improvement resulting from this correction in Fig. 5 is significant.

The scaling behavior with a large momentum cutoff  $\Lambda$  in place is complicated, but it is fairly straightforward to understand it and to find a sequence of approximations that systematically improve the non-perturbative results. The calculations become increasingly complicated, but at each order one must improve the running coupling, or fit it to data, and add higher order irrelevant



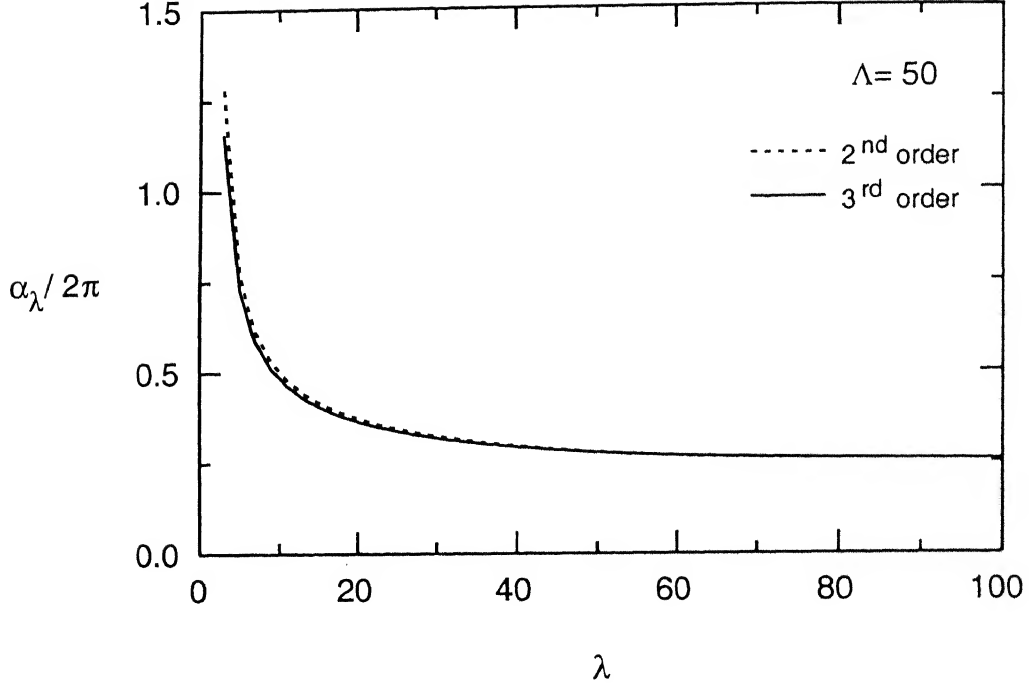


Figure 6: The SRG running coupling for the two-dimensional delta-function potential renormalized to second and third-order obtained with  $\alpha_{\lambda_0=\infty} \rightarrow \alpha_\Lambda = 4\pi/\ln\left(1 + \frac{\Lambda^2}{E_0}\right)$ .

operators. In a field theory we need to let  $\Lambda \rightarrow \infty$  and study the scaling behavior of the theory in the regime where  $\lambda \ll \Lambda$ . Although we do not display a complete set of figures, in Fig. 7 we show what happens to the running coupling as  $\Lambda$  is increased, with the bound state energy fixed at one.

As is evident in the exact solution, as  $\Lambda$  increases the coupling decreases. When  $\lambda \gg \Lambda$ , the coupling runs slowly and stays near its asymptotic value. As  $\lambda$  approaches  $\Lambda$  the coupling begins to run noticeably, and when  $\lambda$  becomes much less than  $\Lambda$  the coupling approaches a universal curve that is insensitive to its asymptotic value. Plots of the error in the binding energy for various approximations and different values of  $\Lambda$  closely resemble Fig. 11, with two scaling regimes whose boundary is  $\lambda = \Lambda$ .

We close this section by reminding the reader that in all of these calculations there is only one free parameter. In a realistic calculation we would fit this parameter to a binding energy and we would expect to see residual errors in other observables that is inversely proportional to powers of the cutoff and logarithms of the cutoff.

## 4 Conclusions

We have illustrated the similarity renormalization group method for producing effective cutoff hamiltonians using the two-dimensional delta-function potential. We have shown that the SRG with coupling coherence leads to errors that scale as inverse powers of the cutoff and inverse logarithms of the cutoff. The SRG with coupling coherence requires the same number of parameters as the underlying ‘fundamental’ theory, but the cost is exponentially increasing algebraic complexity to remove errors that contain inverse powers of logarithms of the cutoff.

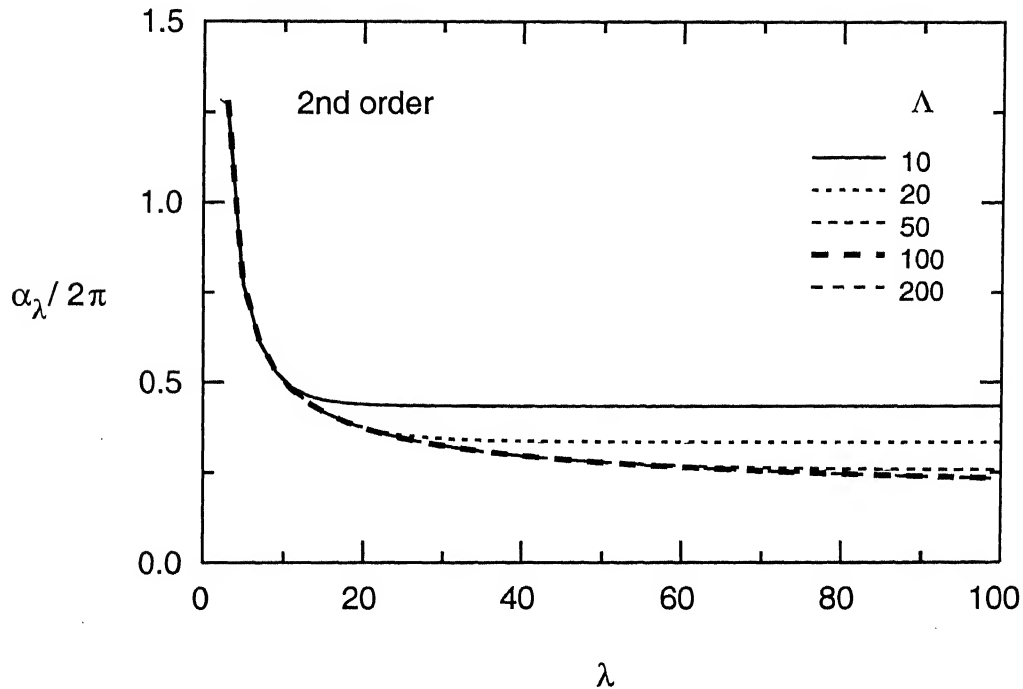


Figure 7: The SRG running coupling for the two-dimensional delta-function potential renormalized to second-order obtained with  $\alpha_{\lambda_0=\infty} \rightarrow \alpha_\Lambda = 4\pi / \ln \left( 1 + \frac{\Lambda^2}{E_0} \right)$ .

## 5 Acknowledgments

We would like to acknowledge many useful discussions with Brent Allen, Martina Brisudova, Dick Furnstahl, Stan Glazek, Billy Jones, Roger Kylin, Rick Mohr, Jim Steele, and Ken Wilson. This work was supported by National Science Foundation grant PHY-9800964, and S.S. was supported by a CNPq-Brazil fellowship (proc. 204790/88-3).

## References

- [1] J. Schwinger, "Quantum Electrodynamics," Dover, New York, 1958.
- [2] P. A. M. Dirac, *Theorie du Positron*, (7-eme Conseil du Physique du Solvay: Structure et propiete de noyaux atomiques, Octobre 1933), pp. 203-230, Gauthier-Villars, Paris, 1934.
- [3] P.A.M. Dirac, in "Perturbative Quantum Chromodynamics," (D.W. Duke and J.F. Owens, Eds.), Am. Inst. Phys., New York, 1981.
- [4] E. C. G. Stueckelberg and A. Peterman, *Helv. Phys. Acta* **26** (1953), 499.
- [5] M. Gell-Mann and F.E. Low, *Phys. Rev.* **95** (1954), 1300.

- [6] L. D. Landau, A. A. Abrikosov and I. M. Khalatnikov, *Doklady* **95** (1954), 497; **96** (1954), 261.
- [7] N. N. Bogoliubov and D.V. Shirkov, *Nuovo Cim.* **3** (1956), 845.
- [8] N. N. Bogoliubov and D.V. Shirkov, "Introduction to the Theory of Quantized Fields," Interscience, New York, 1959.
- [9] K. G. Wilson and J. B. Kogut, *Phys. Rep.* **12C** (1974), 75.
- [10] K. G. Wilson, *Rev. Mod. Phys.* **47** (1975), 773.
- [11] D. J. Gross and F. Wilczek, *Phys. Rev. Lett.* **30** (1973), 1343;  
H. D. Politzer, *Phys. Rev. Lett.* **30** (1973), 1346.
- [12] Ya. B. Zel'dovich, *Soviet Physics JETP* **11** (1960), 594.
- [13] F. A. Berezin and L. D. Faddeev, *Sov. Math. Dokl.* **2** (1961), 372.
- [14] C. Thorn, *Phys. Rev. D* **19** (1979), 639.
- [15] K. Huang, "Quarks, Leptons and Gauge Fields," World Scientific, Singapore, 1982.
- [16] S. Albeverio, F. Gesztesy, R. Hoeg-Krohn and H. Holden, "Solvable Models in Quantum mechanics," Springer-Verlag, New York, 1988.
- [17] C. R. Hagen, *Phys. Rev. Lett.* **64** (1990), 503.
- [18] R. Jackiw, in "M. A. B. Beg Memorial Volume," (A. Ali and P. Hoodbhoy, eds.), World Scientific, Singapore, 1991.
- [19] P. Goddard and R. Tarrach, *Am. J. Phys.* **59** (1991), 70;  
C. Manuel and R. Tarrach, *Phys. Lett. B* **328** (1994), 113.
- [20] J. F. Perez and F. A. B. Coutinho, *Am. J. Phys.* **59** (1991), 52.
- [21] L. R. Mead and J. Godines, *Am. J. Phys.* **59** (1991), 935.
- [22] C. Manuel and R. Tarrach, *Phys. Lett. B* **301** (1994), 72.
- [23] T. J. Fields, K. S. Gupta, and J. P. Vary, *Mod. Phys. Lett. A* **11** (1996), 2233.
- [24] K. S. Gupta and S. G. Rajeev, *Phys. Rev. D* **48** (1993), 5940;  
R. J. Henderson and S. G. Rajeev, *Intl. J. Mod. Phys. A* **10** (1995), 3765;  
R. J. Henderson and S. G. Rajeev, *J. Math. Phys.* **38** (1997), 2171.
- [25] D. K. Park, *J. Math. Phys.* **36** (1995), 5453.
- [26] S. K. Adhikari and T. Frederico, *Phys. Rev. Lett.* **74** (1995), 4572;  
S. K. Adhikari, T. Frederico and I. D. Goldman, *Phys. Rev. Lett.* **74** (1995), 487;  
S. K. Adhikari and A. Ghosh, *J. Phys. A* **30** (1997), 6553;  
C. F. de Araujo, Jr., L. Tomio, S. K. Adhikari and T. Frederico, *J. Phys. A* **30** (1997), 4687.
- [27] R. M. Cavalcanti, quant-ph/9801033 (1998).
- [28] R. J. Henderson and S. G. Rajeev, *J. Math. Phys.* **39** (1998), 749.
- [29] D. R. Phillips, S. R. Beane, and T. D. Cohen, *Ann. Phys. (N.Y.)* **263** (1998), 255.
- [30] S. D. Glazek and K.G. Wilson, *Phys. Rev. D* **48** (1993), 5863.
- [31] S. D. Glazek and K.G. Wilson, *Phys. Rev. D* **49** (1994), 4214.

- [32] F. Wegner, *Ann. Physik* (Berlin) **3** (1994), 77.
- [33] K. G. Wilson, *Phys. Rev.* **140** (1965), B445.
- [34] K. G. Wilson, *Phys. Rev.* **D2** (1970), 1438.
- [35] C. Bloch, *Nucl. Phys.* **6** (1958), 329.
- [36] R. Oehme and W. Zimmermann, *Commun. Math. Phys.* **97** (1985), 569;  
R. Oehme, K. Sibold and W. Zimmermann, *Phys. Lett. B* **147** (1984), 115.
- [37] R. J. Perry and K. G. Wilson, *Nucl. Phys. B* **403** (1993), 587;  
R. J. Perry, *Ann. Phys.* (N.Y.) **232** (1994), 116.
- [38] S. Coleman and E. Weinberg, *Physical Review D* **7** (1973), 1888.

### 3. Quantum Field Theory and the Standard Model: Bird's-eye-view

V.Novikov \*

ITEP, Moscow,Russia and University of Guelph, Guelph,Canada

#### Abstract

We present the panoramic view of the Field Theory ideas that are used in the Standard Model of elementary particles physics.

## 1 Introductory remarks : Brief Review of QFT

The **Standard Model** (SM) pretends to be the Fundamental Theory of Nature. It gives perfect description of the physical phenomena from the scale of binding energy of electrons in molecules and atoms (i.e. from the electron-volt (eV) and less) to the scale of energies of particles in modern accelerators ( i.e. to the hundreds GeV;  $1 \text{ GeV} = 10^9 \text{ eV}$  ). This region of energy has been studied experimentally during this century and it seems there is nothing beyond SM. Thus we believe that the Laws of Nature in this region of energy are known and that the accuracy of the description of any physical phenomenon depends only on our ability to perform mathematical calculations.

On the other hand there still exist a few physical questions that remain not answered in the framework of the SM. Say, we do not understand yet the spectrum of mass of quarks and leptons. Thus many people feel that the SM is incomplete theory and that there should be some New Physics beyond the Standard Model.

The most recent evidence of the great success of the SM was connected with detailed study of properties of Z boson - fundamental particle that mediates electroweak interactions. Special huge electron - positron colliders SLC (at SLAC) and LEP I (at CERN) were constructed at the end of 1980s to measure the parameters of Z boson decays with extremely high accuracy. More than 2000 experimentalists during ten years were involved in these unique experiments. Hundreds of theorists carried out detailed calculations of the required tiny corrections. The result of the collective quest for truth was remarkable - with the accuracy of the order of several thousandths theoretical calculations reproduce the whole set of the experimental data!

The Standard Model is formulated in terms of the renormalizable **Quantum Field Theory** (QFT). Another basic concept that lies in the foundation of the SM is the **Principle of local gauge invariance**. According to this principle the form of gauge interactions is uniquely specified by gauge invariance. The interactions are mediated by vector gauge bosons that are associated with the group of symmetry. The symmetry group of the Standard Model is the product  $SU(3) \times SU(2) \times U(1)$ . Eight  $SU(3)$  gauge bosons (gluons) are responsible for strong interactions. Four  $SU(2) \times U(1)$  gauge bosons ( two W bosons, Z boson and photon) mediate electroweak interaction. In this paper we will consider electroweak interactions only.

In the past Quantum Field Theory was considered as the esoteric theory accessible to the small group of experts. Now QFT provides the working language inside the community of high energy physicists and the basic ideas of QFT, such as Feynman diagrams, are familiar to any member of the community. There are number of excellent textbooks on QFT. (The very incomplete list of the most recent books can be found in the **References**). As a rule they are rather lengthy. To get actual understanding of QFT one has to study one of these books.

---

\*Email : novikov@heron.itep.ru

The goal of this paper is not to provide systematic introduction to QFT or to SM, it is much more modest. We try to give a sort of panoramic view on the basic concepts, notions and relations of QFT and SM without long derivations and boring formalism. Nevertheless the paper is written not for pedestrians but for physicists. We suppose a general background in quantum mechanics and electrodynamics.

The article arose from the lectures at 1998 European School of High-Energy Physics.

## 1.1 Preliminaries: Particles and Fields

First one should answer the very natural question why Field Theory is used to describe Particle Physics.

Indeed in the Classical Physics particles and fields are very different dynamical systems. The system of particles has finite number of degrees of freedom  $N$ . To describe the physical state of particles one has to know the general coordinates  $q_i(t)$  ( $i = 1, 2, \dots, N$ ) and their time derivatives  $\dot{q}_i(t)$  (or conjugate momenta  $p_i(t)$ ) at any time  $t$ . Euler-Lagrange or Hamilton equations of motion govern the dynamics of the system. Either we study the bounded motion or the scattering processes at any time the number of degrees of freedom of the system is fixed.

Field theory is a theory of the system with infinite number of degrees of freedom. The well known example of the fields in Classical Physics is the electromagnetic field. To describe the electromagnetic fields we have to know four-potential  $A_\mu$  at every space point  $x$  and Maxwell equations govern the evolution of the fields  $A_\mu$  in time.

Particles and fields are quite different. This is evident!

In Quantum Mechanics (QM) of non-relativistic particles dynamical system with  $N$  degrees of freedom is described by wave function

$$\Psi(q, t) = \Psi(q_1, \dots, q_N; t) \quad (1.1)$$

that satisfies the wave equation

$$i \frac{\partial}{\partial t} \Psi(q, t) = H(p, q) \Psi(q, t) \quad (1.2)$$

where  $H(p, q)$  is the Hamiltonian. Canonical conjugate coordinates  $p$  and  $q$  are replaced by operators satisfying canonical commutation rules  $[p, q] = -i$ . Thus in coordinate representation the operator of momenta  $p$  is differential operator:  $p = -i\partial/\partial q$ . The number of degrees of freedom  $N$  is supposed to be fixed exactly like in Classical Mechanics.

The first quantization of electromagnetic fields as the dynamical system with infinite number of degrees of freedom had been done in 1926 by Born, Heisenberg and Jordan just in their second paper on QM. They represented radiation electromagnetic field as an infinite set of harmonic oscillators and quantized these oscillators. They found that excitations of the oscillators behave exactly like a free massless particles – photons, but the number of photons was not fixed. Photons could be created and annihilated by charged particles. Quantized theory of electromagnetic field became a theory of particles – photons. Photons were not "bound" inside charged particles, they were created from "nothing" by scattered charged particles. Though the physical idea of photons was not very new (it was introduced by Einstein twenty years before this paper), this step was very important. The formal quantization of electromagnetic field showed that **the quantized field is equivalent to the system of particles that can be created and destroyed**.

Nevertheless for some time physicists continued to treat massive particles (electrons) and electromagnetic fields (photons) as something different. They tried to find a relativistic version of the Schroedinger wave equation (1.2) for the particles at high energy. The first such equation for spin 0 relativistic particles (Klein-Gordon equation) was written in 1926 by many authors

$$-\partial_\mu \partial_\mu \Phi(x) = m^2 \Phi(x) \quad (1.3)$$

where  $\partial_\mu = \partial/\partial x_\mu$  and  $\Phi(x)$  is a complex function of  $x = (t, \vec{x})$ ,  $m$  is a mass of particle.

Immediately it was pointed out that eq. (1.3) and the function  $\Phi(x)$  can't be interpreted as the wave equation and the wave function. Such interpretation led to a number of physical paradoxes. Later in 1928 Dirac suggested another relativistic equation (for spin 1/2 particles):

$$(i\gamma_\mu\partial_\mu - m)\Psi(x) = 0 \quad (1.4)$$

Here  $\Psi$  is a column with 4 complex components (4-spinor) and  $\gamma_\mu$  are  $4 \times 4$  matrices.

The troubles with interpretation of eq. (1.4) as the one-particle relativistic wave equation were not so evident as for the case of eq. (1.3). But the truth is that for any relativistic processes the single particle description should break down. Indeed any relativistic system has infinite numbers of degrees of freedom. The more energy we pump into the system, the more degrees of freedom can be excited. Say any process in Quantum Electrodynamics (QED) can be accompanied by creation of any number of additional  $e^+e^-$  pairs. These pairs are not hidden inside initial particles, they are created from vacuum.

Like in the case of photons the natural description of relativistic system with varied number of degrees of freedom is the quantum theory of the appropriate field. It is wrong to divide world on particles and fields - one has to use the quantum field theory for everything. QFT is the right language for dealing with particle physics!

From this point of view the description of QFT as a second quantization sounds very misleading. Nobody quantizes wave functions since both Klein-Gordon and Dirac equations (as well as Maxwell equations) are not relativistic equations for wave function. They are field equations for scalar and for spinor fields respectively. Moreover nobody quantizes classical fields as well. There is no straight way from Classical Physics to Quantum Theory. The Fundamental Theory is the Quantum Field Theory. The Classical Theory is the special limit of QFT and one should start from QFT and not vice versa.

## 1.2 Quantization of Free Fields and Fock space.

In this subsection we demonstrate that quantum field theory indeed describes particles.

Consider first the very simple example - free scalar field:

$$\Phi(x) = \Phi(\vec{x}, t) \quad (1.5)$$

In the "Classical Theory"  $\Phi(x)$  is a real function of space-time point  $x_\mu = (t, \vec{x})$ . It represents the set of general coordinates of the system with label  $\vec{x}$ . For free field there is only one choice for Lorenz-invariant Lagrangian density  $\mathcal{L}(\Phi, \partial_\mu \Phi)$  :

$$\mathcal{L}(\Phi, \partial_\mu \Phi) = \frac{1}{2} \{ \partial_\mu \Phi \partial_\mu \Phi - m^2 \Phi^2 \} \quad (1.6)$$

where  $\partial_\mu \Phi = \frac{\partial}{\partial x_\mu} \Phi$  and the coefficient  $m$  has dimension of the mass.

The action  $S$  is given by

$$S = \int d^4x \mathcal{L}(\Phi, \partial_\mu \Phi) \quad (1.7)$$

To quantize this field we need Hamiltonian and canonically conjugate momenta. The Hamiltonian density  $\mathcal{H}$  is constructed according to the rules of Hamiltonian dynamics

$$\mathcal{H} = \pi \frac{\delta \mathcal{L}}{\delta \Phi} - \mathcal{L} = \frac{1}{2} \{ \pi^2 + (\nabla \Phi)^2 + m^2 \Phi^2 \} \quad (1.8)$$

where

$$\pi = \frac{\delta \mathcal{L}}{\delta \dot{\Phi}} = \dot{\Phi} = \partial_0 \Phi$$

is the conjugate momenta.

We use the natural units where  $c \equiv 1$  and  $\hbar \equiv 1$ . In these units the action is dimensionless

$$[S] = m^0 ,$$

and for dimension of the other quantities one gets

$$\begin{aligned} [E] &= [p] = m \\ [x] &= m^{-1} \\ [\mathcal{L}] &= [\mathcal{H}] = m^4 \\ [\phi] &= m \end{aligned} \tag{1.9}$$

Euler-Lagrange equations of motion are derived from the Hamilton least action principle

$$\delta S = 0 \tag{1.10}$$

and coincide with Klein-Gordon equation

$$(\partial^2 + m^2)\Phi = 0 \tag{1.11}$$

If we consider the plane wave solution for eq. (1.11)

$$\Phi_{\vec{p}}(x, t) = a(p, t)e^{i\vec{p}\vec{x}} \tag{1.12}$$

the equation for the amplitude  $a$

$$\ddot{a} + (\vec{p}^2 + m^2)a = 0 \tag{1.13}$$

looks exactly like equation for linear oscillator with frequency

$$\omega^2(p) = \vec{p}^2 + m^2 \ .$$

It is crucial that the dependence of frequency  $\omega$  on  $\vec{p}$  is exactly the same as the dependence of particle energy on momentum  $\vec{p}$  (in the units  $\hbar = c = 1$ ). This is why one can use free fields to describe free particles.

The general solution in the periodic box can be presented as a superposition of the solutions (1.12)

$$\Phi(x) = \sum_p [a(p)e^{-ipx} + a^+(p)e^{ipx}] \tag{1.15}$$

where

$$\begin{aligned} px &= p_\mu x_\mu = p_0 x_0 - \vec{p}\vec{x} \\ p_0 &= \omega(p) = \sqrt{\vec{p}^2 + m^2} \\ \sum_p &= \int \frac{d^3 p}{(2\pi)^3 2p_0} \end{aligned}$$

In terms of these variables the Hamiltonian is equal to

$$H = \int d^3 x \mathcal{H} = \sum_p \frac{1}{2} \omega(p) [aa^+ + a^+ a] \tag{1.16}$$

This is the Hamiltonian for the set of decoupled linear oscillators. In the Classical theory coefficients  $a(p)$  and  $ia^+(p)$  are canonically conjugate variables. According to canonical procedure of quantization we have to replace them by operators that satisfy commutation relations

$$[a(\vec{p}), a^+(\vec{p}')] = \delta_{\vec{p}\vec{p}'} \tag{1.17}$$

$$[a(\vec{p}), a(\vec{p}')] = [a^+(\vec{p}), a^+(\vec{p}')] = 0$$

Operators  $a(p)$  and  $a^+(p)$  are familiar from QM. They are the annihilation and creation operator for oscillator with frequency  $\omega(\vec{p})$ .

The **Fock space** is the Hilbert space of the states with definite values of the operators of particle number:  $N(p) = a^+(p)a(p)$ :



vacuum state

$$\begin{cases} |0\rangle \\ a(p)|0\rangle \equiv 0 \end{cases},$$

one-particle states

$$|p\rangle = a^\dagger(p)|0\rangle \quad (1.18)$$

two-particle states

$$\begin{aligned} |p_1, p_2\rangle &= a^\dagger(p_1)a^\dagger(p_2)|0\rangle \\ |p; p\rangle &= \sqrt{2}a^\dagger(p)a^\dagger(p)|0\rangle \end{aligned}$$

etc.

Commutation relation eq. (1.17) corresponds to Bose-Einstein statistics for spin 0 particle. Indeed

$$|p_1, p_2\rangle = +|p_2, p_1\rangle$$

The operator of energy

$$H = \sum \omega(p) \left[ N(p) + \frac{1}{2} \right] \quad (1.20)$$

is well defined operator and it is bounded from below by the vacuum energy

$$E_{vac} = \sum \frac{1}{2} \omega(p) \quad (1.21)$$

The space of excited states of field oscillators - **Fock space** represents the **states of free particles** with mass  $m$ , with given momenta, and with positive energy  $p_0 = \omega(p) = \sqrt{p^2 + m^2}$ .

The theory of the complex scalar fields  $\Phi(x) = \frac{1}{\sqrt{2}}(\Phi_1 + i\Phi_2)$  with Lagrangian density

$$\mathcal{L} = (\partial_\mu \Phi^\dagger \partial_\mu \Phi) - m^2 \Phi^\dagger \Phi \quad (1.22)$$

is equivalent to the theory of two different scalar particles with degenerate masses. The general solution of the field equations can be presented in the form

$$\Phi(x) = \sum_p (a(p)e^{-ipx} + b(p)^\dagger e^{ipx}) \quad (1.23)$$

where the operators  $(a, a^\dagger)$  and  $(b, b^\dagger)$  are creation and annihilation operators for two different particles with the same masses but with the opposite electric charges (see the next chapter). This is some new phenomena:

**QFT predicts that for any particle there should exist anti-particle.**

For Dirac spinor field  $\Psi(x)$  the Lagrangian density can be written as

$$\mathcal{L} = \bar{\Psi}[i\gamma_\mu \partial_\mu - m]\Psi \quad (1.24)$$

The dimension of field  $\Psi$ :  $[\Psi] = m^{3/2}$ . The plane wave solutions of the Dirac equation look like

$$u(p, \lambda)e^{ipx}, \quad (\lambda = \pm 1/2) \quad (1.25)$$

$$v(p, \lambda)e^{-ipx}, \quad (\lambda = \pm 1/2)$$

where  $u(p, \lambda)$ ,  $v(p, \lambda)$  satisfy equations

$$(\gamma_\mu p_\mu - m)u(p, \lambda) = 0$$

$$(1.26)$$

$$(\gamma_\mu p_\mu + m)v(p, \lambda) = 0$$

and  $\lambda = \pm 1/2$  label the independent solution with different value of the spin projection on momenta  $\vec{p}$ . The general solution of Dirac equation can be presented as the superposition of plane wave solutions

$$\Psi(x) = \sum_{\vec{p}, \lambda} \{a(p, \lambda)u(p, \lambda)e^{-ipx} + b^+(p, \lambda)v(p, \lambda)e^{ipx}\}. \quad (1.27)$$

The dynamical coordinates  $a(p, \lambda)$ ,  $ia^+(p, \lambda)$  and  $b(p, \lambda)$ ,  $ib^+(p, \lambda)$  are conjugate variables.

The next step is the quantization. We have to consider  $a(p, \lambda)$  and  $b(p, \lambda)$  as the operators in the Fock space. The great surprise is that to get well defined operator of energy we should not follow the procedure of canonical quantization. Instead the operators  $a(p, \lambda)$  and  $b(p, \lambda)$  should satisfy anti-commutation (not commutation) relations

$$\{a(p, \lambda), a^+(p', \lambda')\} = \{b(p, \lambda), b^+(p', \lambda')\} = \delta_{pp'} \delta_{\lambda\lambda'} \quad (1.28)$$

$$\{a, a\} = \{a^+, a^+\} = \{b, b\} = \{b^+, b^+\} = 0$$

with  $\{A, B\} = AB + BA$ .

Only in this way we get that the energy is bounded from below and that the local observables at equal times commute for separated space points. Anticommutators imply Fermi-Dirac statistic for spin 1/2 particle. The operators  $a(p, \lambda)$ ,  $a^+(p, \lambda)$  and  $b(p, \lambda)$ ,  $b^+(p, \lambda)$  are annihilation and creation operators for particles and anti-particles respectively.

These two examples demonstrate the famous spin-statistics theorem:

**QFT can be self-consistent if and only if the identical particles with integer spin obey Bose-Einstein statistics and the particles with half-integer spin obey Fermi-Dirac statistics.**

For Electromagnetic Field  $A_\mu(x)$  the Lagrangian density is

$$\begin{aligned} \mathcal{L} &= -\frac{1}{4}F_{\mu\nu}F_{\mu\nu} + ej_\mu^{em}A_\mu, \\ F_{\mu\nu} &= \partial_\mu A_\nu - \partial_\nu A_\mu, \\ [A_\mu] &= m, [j_\mu] = m^3. \end{aligned} \quad (1.29)$$

The general solution for radiation field can be presented as a superposition of plane waves

$$A_\mu(x) = \sum_{\vec{p}, \lambda} \{a(p, \lambda)\varepsilon_\mu(p, \lambda)e^{-ipx} + a^+(p, \lambda)\varepsilon_\mu^*(p, \lambda)e^{ipx}\}.$$

Because of the gauge invariance the formal quantization of electromagnetic field technically is a little bit more subtle matter. We will not touch this subject in our paper.

We have constructed the physical states for free particles. It is also useful to have the amplitudes that describe the propagation of free particles from one space-time point to another one. Consider scalar fields first. The part of the field operator (1.15) that represents the terms with positive frequencies

$$\Phi^{(+)} = \sum_{\vec{p}} a(\vec{p})e^{-ipx}$$

annihilates particle at point  $x$ , while the operator

$$\Phi^{(-)} = \sum_{\vec{p}} a^+(\vec{p})e^{ipx} \quad (1.30)$$

creates the particle at point  $x$ .

Thus the vacuum expectation of the time-ordered product of fields

$$D_F(x, 0) = \langle 0 | T \{ \Phi(x) \Phi(0) \} | 0 \rangle \quad (1.31)$$

represents the amplitude for a particle to propagate from point 0 to point  $x$  (Feynman propagator). Time ordering implies that creation always comes before annihilation. Here the Dyson's time-ordered product of operators  $T$  is defined as follows:

$$T \{ \Phi(x), \Phi(0) \} = \Theta(x_0) \Phi(x) \Phi(0) + \Theta(-x_0) \Phi(0) \Phi(x) \quad (1.32)$$

where the step function is equal to

$$\Theta(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x < 0 \end{cases}$$

The Feynman propagator of scalar particle has very simple form in the momentum representation

$$D_F(p) = \frac{i}{p^2 - m^2 + i\epsilon} \quad (1.33)$$

Feynman propagator of electrons  $S_F(x)$

$$S_F = \langle 0 | T \{ \Psi(x) \bar{\Psi}(y) \} | 0 \rangle \quad (1.34)$$

$$T \{ \Psi(x), \bar{\Psi}(0) \} = \Theta(x_0) \Psi(x) \bar{\Psi}(0) - \Theta(-x_0) \bar{\Psi}(0) \Psi(x).$$

in momentum representation looks like

$$S_F(p) = \frac{i}{\not{p} - m} \quad (1.35)$$

where  $\not{p} = \gamma_\mu p_\mu$ .

The photon propagator in the so called Feynman gauge looks like

$$D_F^{\mu\nu} = \frac{(-i)g_{\mu\nu}}{p^2 + i\epsilon} \quad (1.36)$$

That is the end of our excursion into the quantization of the free fields.

### 1.3 Interaction: Feynman Diagrams.

What we do understand well is the QFT in the framework of perturbation theory when one can separate Lagrangian into quadratic free term  $\mathcal{L}_0$  and interaction term  $\mathcal{L}_{int}$  that can be considered as a small perturbation to  $\mathcal{L}_0$ . Free field theory provides the asymptotic  $|in\rangle$  and  $|out\rangle$  states for particles (and anti-particles). The nonlinear interaction term  $\mathcal{L}_{int}$  in perturbation theory provides the transition amplitudes from one asymptotic state to another one.

Transitions from the initial asymptotic states to the final states are described by means of unitary  $S$ -matrix:  $S^\dagger S = I$

$$\langle f | S | i \rangle = \langle f | i \rangle + (2\pi)^4 i \delta^{(4)}(\Sigma p_f - \Sigma p_i) \langle f | T | i \rangle \quad (1.37)$$

where  $i$  and  $f$  refer to initial and final state.

In perturbation theory  $S$  matrix is given by Feynman-Dyson series of integrals over time-ordered products of  $\mathcal{L}_{int}$  in the so called interaction representation  $\mathcal{L}_I$

$$S = T \exp \{ i \int d^4x \mathcal{L}_I \} = \quad (1.38)$$

$$= I + i \int d^4x \mathcal{L}_I(x) + \frac{i^2}{2} T \left\{ \int d^4x_1 \mathcal{L}_I(x_1); \int d^4x_2 \mathcal{L}_I(x_2) \right\} + \dots$$

This representation of  $S$ -matrix can be translated into the language of Feynman diagrams. According to Feynman there is a set of two basic elements: propagators and vertices. Propagators were found in the previous subsection, vertices depend on interaction. To calculate the amplitude of any of physical processes one has

- 1) to draw all distinct diagrams for the process combining propagators and vertices in all possible ways,
- 2) to assign amplitudes for the propagators and for the vertices in given diagram and to multiply them,
- 3) to sum the contribution of all distinct diagrams.

Consider as an example the QED of leptons. In this case the interaction is given by the product of electromagnetic current  $j_\mu^{em}(x)$  and 4-potential  $A_\mu(x)$

$$\mathcal{L}_{int} = j_\mu^{em}(x) A_\mu(x) \quad (1.39)$$

$$j_\mu^{em}(x) = (-ie) \{ \bar{e}(x) \gamma_\mu e(x) + \dots \},$$

where  $e(x)$  represents electron field operators, dots are for muon and  $\tau$  leptons contribution, and  $e$  is electric charge of the proton.

Feynman rules for this QED Lagrangian are summarized in Fig. 1.

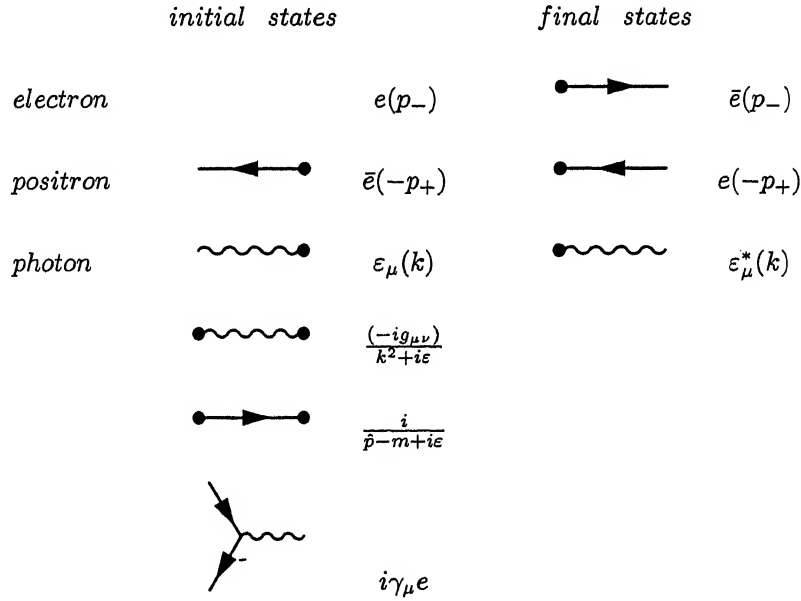
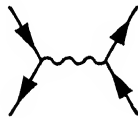


Figure 1: Feynman rules of QED.

Using these rules one can easily construct any transition amplitudes.

Consider as an example the process  $e^+e^- \rightarrow \mu^+\mu^-$ . There is one diagram for this process



The amplitude  $T$  is equal

$$iT(e^+e^- \rightarrow \mu^+\mu^-) = (-ie)^2 j_\alpha^{(e)} \frac{(-ig_{\alpha\beta})}{q^2} j_\beta^{(\mu)} \quad (1.40)$$

where

$$j_\alpha^{(e)} = \bar{v}(-p_+) \gamma_\alpha v(p_-)$$

$$j_\beta^{(\mu)} = \bar{\mu}(k_-) \gamma_\beta \mu(-k_+)$$

This is the example of the amplitude in the lowest order in the coupling constant  $e$ . It contains no loops. There is special name for such diagrams - tree diagrams.

The steps from the QFT to the calculation of cross sections and of the decay rates are very simple. Using Feynman diagrams one calculates first the amplitudes of the process  $T$ , then the square of the modulus  $T$  and finally one performs the summation over all degenerate final states: Probability  $\sim \Sigma |T|^2$ .

More precisely the cross sections are calculated by the formula

$$d\sigma_{fi} = \frac{1}{2\sqrt{\lambda(s, m_1^2, m_2^2)}} |T_{fi}|^2 d\tau$$

where

$$d\tau = (2\pi)^4 \delta^4(p_+ + p_- - \Sigma p_f) \prod_{j=1}^N \frac{d^3 p_j}{(2\pi)^3 2E_j}$$

is N-particle phase space and

$$\lambda(s, m_1^2, m_2^2) = 4[(p_1 p_2)^2 - m_1^2 m_2^2]$$

is relativistic flux.

The decay rates are given by formula

$$d\Gamma = \frac{1}{2E} |T|^2 d\tau$$

There exists well developed routine technology of that type of calculations.

Consider now one-loop diagrams. We focus on the corrections to the photon propagator with momentum  $q$ . There is only one such diagrams



According to Feynman rules the correction  $\delta D_{\mu\nu}$  to the propagator is equal to

$$\delta D_{\mu\nu} = \frac{(-ig_{\mu\alpha})}{q^2} (-i\Pi_{\alpha\beta}(q)) \frac{(-i)g_{\beta\nu}}{q^2} \quad (1.41)$$

where

$$(-i)\Pi_{\alpha\beta}(q) = e^2 \int \frac{d^4 p}{(2\pi)^4} (-1) \text{Tr} \gamma_\alpha \frac{1}{\hat{p} - m + i\varepsilon} \gamma_\beta \frac{1}{\hat{p} - \hat{q} - m + i\varepsilon} \quad (1.42)$$

For large virtual momenta  $p$  the loop correction diverges quadratically. If we regularize integral by means of introduction cut-off  $\Lambda$  the result of integration looks like

$$\Pi_{\alpha\beta} \simeq e^2 g_{\alpha\beta} \int_0^\Lambda \frac{d^4 p}{p^2} \simeq g_{\alpha\beta} e^2 \Lambda^2 \rightarrow \infty \quad (1.43)$$

for  $\Lambda \rightarrow \infty$ .

More sophisticated regularization demonstrates that quadratic divergence actually disappears and that integral diverges logarithmically.

This is the simplest example of the problem of divergences in QFT. It was a great success of theoretical physics when Dyson, Feynman, Schwinger and Tomonaga in the late 40th explained how to work with such theories.

## 1.4 Renormalizable Field Theories QED.

The general philosophy of renormalization can be formulated in the following way:

1) Suppose that we can separate all quantum fluctuations into the "fast" fluctuations (i.e. with virtual momenta  $p > \Lambda$ ) and into the "slow" ones ( $p < \Lambda$ ), where  $\Lambda$  is arbitrary large parameter (cut-off).

2) Suppose that we can integrate over the "fast" fluctuations in some way (even though the physics at small distances ( $p > \Lambda$ ) can be unknown yet).

3) For "slow" fluctuations we get "effective field theory" with  $\mathcal{L}^{eff}(\Lambda)$  or  $S^{eff}(\Lambda)$  that govern the dynamics at low momenta. Effective field theory parameters depend on cut-off  $\Lambda$ .

4) For renormalizable theories  $S^{eff}(\Lambda)$  depends on finite number parameters and interaction terms. In this case one can express these parameters in terms of the same number of low-energy parameters and cut-off  $\Lambda$ . The cut-off  $\Lambda$  can be rid of low-energy observables if one rewrites them in terms of low-energy parameters.

The renormalizable quantum field theories are the very reasonable from the physical point of view. They describe the situation when the large scale dynamics does not depend on the details of short distance physics.

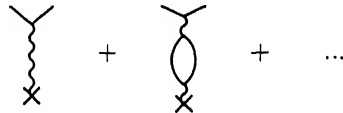
Let us look how this program works in the case of QED. The form of the effective Lagrangian is fixed by gauge invariance (see next chapter)

$$\begin{aligned} \mathcal{L}(\Lambda) = & -\frac{1}{4}(F_{\mu\nu}^B)^2 + \bar{\Psi}_B(i\gamma_\mu\partial_\mu - m_B)\Psi_B - \\ & -e_B\bar{\Psi}_B\gamma_\mu\Psi_B A_\mu^B \end{aligned} \quad (1.46)$$

where all quantities with label  $B$  depend on  $\Lambda$ .

Consider the scattering of heavy charged particle on the Coulomb field.

In this case we have to sum up the loop corrections to the photon propagator with electron-positron pairs inside



As a result the amplitude of Coulomb scattering can be written as

$$T = \frac{e_B^2(\Lambda)}{1 - \frac{e_B^2(\Lambda)}{12\pi^2} \ln \frac{\Lambda^2}{m_e^2}} \cdot \frac{1}{q^2} \quad (1.47)$$

The coefficient in front of  $1/q^2$  is by definition the charge of particle ( $1/q^2$  corresponds to  $1/r$  dependence in the Coulomb law). So we claim that combination

$$e_{ph}^2 = \frac{e_B^2(\Lambda)}{1 - \frac{e_B^2(\Lambda)}{42\pi^2} \ln \frac{\Lambda^2}{m_e^2}} \quad (1.48)$$

is the physical charge that can be experimentally measured. In this way we find  $e_B^2(\Lambda)$  as a function of physical charge and cut-off  $\Lambda$ .

In the similar way one can define the physical electron mass  $m_{ph} = m_e$  as a pole in the exact propagator of the electron.

Now we are able to formulate the main statement: if one rewrites the amplitudes of any QED process that depend on  $e_B$ ,  $m_B$  and  $\Lambda$  in terms of  $e_{ph}$ ,  $m_{ph}$  the dependence on  $\Lambda$  in these amplitudes will disappear for large  $\Lambda$ !

## 1.5 Non-renormalizable Theories.

The first theory of weak interactions was formulated by Fermi in 1934. It was very similar to QED. The Lagrangian of interaction was equal to a product of two vector currents (after the discovery of  $P$  and  $C$  parity violation this 4-fermion theory was modified slightly)

$$\mathcal{L}_W = G_F j_\alpha j_\alpha \quad (1.49)$$

where  $j_\alpha = \bar{\nu}_e \gamma_\alpha e + \dots$

The Fermi coupling constant has dimension -2. Indeed

$$[j] = m^3 ; \quad [\mathcal{L}] = m^4$$

so that

$$[G_F] = m^{-2}.$$

The radiative corrections to 4-fermion interaction are given by diagrams

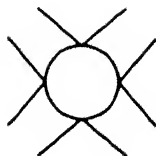


From the dimensional analysis it is clear that the result of calculation should be of the order of

$$G_F (1 + \Sigma(G_F \Lambda^2)^n) (j)^2 \quad (1.50)$$

where  $\Lambda$  is cut-off.

It is also clear that 4-fermion interaction can generate multi-fermion interaction with divergent coupling constant, e.g. 8-fermion interaction



$$\Delta\mathcal{L} = CG_F^4[\ln\Lambda^2 + \Sigma(G_F\Lambda^2)^n](j)^4 \quad (1.51)$$

etc.

In this way we find that  $\mathcal{L}^{eff}$  should have  $j^4$  interaction. Actually it has to have infinite number of terms. Why these divergent corrections still allow one to rely on the lowest order approximation, remained a mystery up to discovery of the SM.

Another example of non-renormalizable theory is the theory of massive vector bosons. Though coupling constant in this case has dimension zero, the longitudinal components of vector fields interact strongly and in general case the theory is non-renormalizable. This is why all naive attempts to construct a renormalizable weak interaction theory with intermediate vector bosons failed in the past.

These are two examples of non-renormalizable theory. In such theories one has to fix infinite number of terms in  $\mathcal{L}^{eff}(\Lambda)$  at the scale  $\Lambda$  (i.e. at small distances  $x \sim \Lambda^{-1}$ ) to reconstruct the amplitudes at low energy. It means that we need a lot of information how Nature is constructed at the scale  $\Lambda$ . Nobody knows yet how to work with non-renormalizable theories.

## 2 Symmetry.

In the Standard Model the principle of local gauge invariance plays the same role as the principle of local Lorentz-invariance plays in General Relativity. This geometrical principle uniquely determines the self-interaction of gauge bosons and their interaction with matter. The great problem is that the same principle prohibits gauge bosons to have masses. As we know massless vector bosons do not exist in Nature with one exception - photon. How to make gauge bosons massive and not to destroy geometrical beauty? The solution of this problem was found in the spontaneous breaking of local symmetry - very beautiful phenomenon discovered first in the solid state physics.

### 2.1 Global symmetry.

**$U(1)$  symmetry.** Consider as an example the theory of free electrons. Electrons are described by Lagrangian density

$$\mathcal{L} = \bar{\psi}(x)(i\gamma_\mu\partial_\mu - m)\psi(x) \quad (2.1)$$

where  $\psi_i (i = 1, 2, 3, 4)$  is 4-component complex field,  $\bar{\psi} = \psi^\dagger\gamma_0$  and  $\gamma_\mu$  are  $4 \times 4$  Dirac matrices, and  $m$  is the electron mass.

The  $U(1)$  global phase transformations

$$\begin{aligned} \psi(x) &\rightarrow \psi'(x) = e^{i\alpha}\psi(x) \\ \bar{\psi}(x) &\rightarrow \bar{\psi}'(x) = \bar{\psi}(x)e^{-i\alpha} \end{aligned} \quad (2.2)$$

leave Lagrangian (2.1) invariant. The global symmetry means that the phase of the transformation is the same at any space-time points  $x$ .

**$SU(2)$  symmetry.** Consider the theory of two complex self-interacting scalar fields with the degenerate masses

$$\mathcal{L} = \partial_\mu\Phi^\dagger\partial_\mu\Phi - m^2\Phi^\dagger\Phi - \frac{\lambda}{4}(\Phi^\dagger\Phi)^2, \quad (2.3)$$

where  $\Phi$  is the two component column (doublet)

$$\Phi = \begin{pmatrix} \varphi^+(x) \\ \varphi^0(x) \end{pmatrix} \quad (2.4)$$

The Lagrangian (2.3) is invariant under global  $SU(2)$  rotations of the complex doublet  $\Phi$

$$\begin{aligned} \Phi(x) &\rightarrow \Phi'(x) = S\Phi(x) \\ \Phi^\dagger(x) &\rightarrow (\Phi'(x))^\dagger = \Phi^\dagger(x)S^\dagger \end{aligned} \quad (2.5)$$



where  $S$  is unitary  $2 \times 2$  matrix

$$\begin{aligned} S^\dagger S &= I \\ \det S &= 1. \end{aligned} \quad (2.6)$$

The  $SU(2)$  transformations are global, i.e. matrix  $S$  does not depend on space-time points  $x$ .

According to the Noether's theorem for any continuous global symmetry of the Lagrangian one can construct the conserved vector currents. This dynamical statement is very beautiful and rather non-trivial. We prove the theorem in the classical field theory.

Let Lagrangian  $L$  depends on the set of fields  $\varphi^i$  and its first derivatives  $\varphi_{,\mu}^i = \partial_\mu \varphi^i$ . For infinitesimal global transformations the variations of fields are equal to

$$\begin{aligned} \delta \varphi^i &= i\epsilon^{(a)} T_{ij}^{(a)} \varphi^j \\ \delta \varphi_{,\mu}^i &= i\epsilon^{(a)} \partial_\mu (T_{ij}^{(a)} \varphi^j) \end{aligned} \quad (2.7)$$

here  $\epsilon^{(a)}$  are the real infinitesimal parameters, one for each independent symmetry transformations, matrices  $T_{ij}^a$  are the generators of the group of transformations in given representations.

The invariance means that the action  $S$  is not changed under transformation (2.7):

$$\delta S = \int d^4x \delta L \equiv 0 \quad (2.8)$$

Calculate now the variation of Lagrangian density directly

$$\begin{aligned} \delta L &= \frac{\partial L}{\partial \varphi^i} \delta \varphi^i + \frac{\partial L}{\partial \varphi_{,\mu}^i} \delta \varphi_{,\mu}^i = \\ &= [\partial_\mu \frac{\partial L}{\partial \varphi_{,\mu}^i}] \delta \varphi^i + \frac{\partial L}{\partial \varphi_{,\mu}^i} \delta \varphi_{,\mu}^i \end{aligned} \quad (2.9)$$

where we have used the Lagrangian equation of motion for  $\varphi^i(x)$

$$\frac{\partial L}{\partial \varphi^i} = \partial_\mu \frac{\partial L}{\partial \varphi_{,\mu}^i} \quad (2.10)$$

Substituting the variations for  $\delta \varphi$  and  $\delta \varphi_{,\mu}$  we get

$$\delta S = i\epsilon^{(a)} \int d^4x \partial_\mu j_\mu^{(a)}(x) = 0, \quad (2.11)$$

where

$$j_\mu^{(a)} = \frac{\partial L}{\partial \varphi_{,\mu}^i} T_{ij}^{(a)} \varphi^j \quad (2.12)$$

Thus we get the conservation of Noether currents

$$\partial_\mu j_\mu^{(a)}(x) = 0 \quad (2.13)$$

and the conservation of the corresponding charges

$$\frac{d}{dt} Q^{(a)}(t) = 0 \quad (2.14)$$

$$Q^{(a)}(t) = \int d^3x j_0^{(a)}(x) \quad (2.15)$$

The generalization of this proof to the Quantum Field Theory requires more advanced techniques such as operators algebra, commutators etc. The final result, i.e. the expression for conserved Noethers current, remains the same. Noethers currents that correspond to  $U(1)$  and  $SU(2)$  symmetries look like follows :

$$\begin{aligned} U(1) &: j_\mu = \bar{\psi} \gamma_\mu \psi, \\ SU(2) &: j_\mu^a = \Phi^\dagger \tau^a \overleftrightarrow{\partial}_\mu \Phi \end{aligned} \quad (2.16)$$

where

$$\Phi^\dagger \overleftrightarrow{\partial}_\mu \Phi = \Phi^\dagger \partial_\mu \Phi - (\partial_\mu \Phi^\dagger) \Phi,$$

and  $\tau^a$  are Pauli matrices.

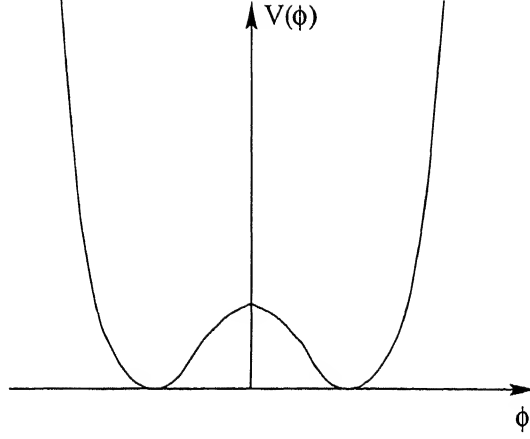


Figure 1: Higgs Potential in Standard Model

## 2.2 Spontaneous Breaking of Global Symmetry : Goldstone Phenomenon

The phenomenon of spontaneous breaking of symmetry is known from the everyday life. Consider, for example, the piece of ferromagnetic material. The interaction of the elementary magnetic moments of electrons inside ferromagnetic is  $O(3)$  invariant. On the other hand at low temperature  $T < T_c$  the total magnetic moment  $\vec{M}$  of the ferromagnetic piece is nonzero. This spontaneous moment breaks  $O(3)$  invariance of the system: ground state is only  $O(2)$  invariant with respect to the rotations around  $\vec{M}$ .

From the solid state physics it is known that the “violated” symmetries are realized as a massless excitations. In field theory analogous phenomenon is known as Nambu-Goldstone realization of symmetry. We consider this phenomenon using a very simple field model studied many years ago by Goldstone.

Consider the theory of complex scalar field  $\varphi(x)$  with Lagrangian

$$\mathcal{L} = \partial_\mu \varphi^\dagger \partial_\mu \varphi - V(|\varphi|^2) \quad (2.17)$$

and with a special choice of potential (see Fig.1)

$$V(|\varphi|^2) = \lambda(|\varphi|^2 - \frac{\eta^2}{2})^2 \quad (2.18)$$

Lagrangian (2.17) is invariant under  $U(1)$  transformations

$$\varphi(x) \rightarrow \varphi'(x) = e^{i\Lambda} \varphi(x) \quad (2.19)$$

and the Noether current is

$$j_\mu = i\varphi^\dagger \overleftrightarrow{\partial}_\mu \varphi \quad (2.20)$$

There are continuously many minima of the potential  $V$

$$\varphi = \frac{1}{\sqrt{2}} \eta e^{i\alpha} \quad (2.21)$$

The vacuum corresponds to one of these minima. This is spontaneous breaking of symmetry: we have chosen as a vacuum state one of the infinite set of minima.

Let the vacuum state corresponds to zero phase  $\alpha = 0$  :

$$\varphi_{cl} = \frac{1}{\sqrt{2}} \eta \quad (2.22)$$

and consider the small fluctuation of fields near this vacuum configuration

$$\varphi = \frac{1}{\sqrt{2}}[\eta + \rho(x) + i\sigma(x)] \quad (2.23)$$

In terms of this fluctuations the potential can be rewritten as

$$V(\varphi) = V(\rho, \sigma) = \frac{\lambda}{2} \{(\sigma^2 + \rho^2)^2 + 4\eta\rho(\rho^2 + \sigma^2) + 4\eta^2\rho^2\} \quad (2.24)$$

The coefficients in front of bilinear terms determine the mass of the fields. So we get a theory of two particles with masses

$$\begin{aligned} M_\rho^2 &= 4\lambda\eta^2 \\ M_\sigma^2 &\equiv 0 \end{aligned} \quad (2.25)$$

We can use more elegant and transparent representation for  $\varphi(x)$  to demonstrate the same phenomenon. Rewrite  $\varphi(x)$  in terms of modulus and phase

$$\varphi(x) = \rho(x)e^{i\sigma(x)} \quad (2.26)$$

In this case

$$\mathcal{L}(\rho, \sigma) = (\partial_\mu \rho)^2 - V(\rho^2) + \rho^2 (\partial_\mu \sigma)^2 \quad (2.27)$$

There is no dependence on the field  $\sigma$  in the potential and therefore this field corresponds to massless particle.

Excitations that corresponds to the motion along the valley of minima are massless! This is **Goldstone phenomenon**.

In Quantum Field Theory there are two ways of realization of symmetry:

- 1) Vacuum state has the symmetry of the action  $S$ . Excitation states are degenerate.
- 2) Vacuum state has lower symmetry than action  $S$ . There are flat direction in configuration space of fields. The motions along these flat directions correspond to massless **Goldstone** particles.

### 2.3 Local U(1) gauge symmetry.

Now we turn to the local gauge symmetries and start with the theory of complex field  $\phi(x)$  described by the (eq.(2.17))

$$\mathcal{L} = \partial_\mu \phi^\dagger \partial_\mu \phi - V(\phi^\dagger \phi).$$

This Lagrangian is invariant under global  $U(1)$  transformation

$$\phi(x) \rightarrow \phi'(x) = e^{i\Lambda} \phi(x).$$

Consider now the local  $U(1)$  transformation when we change the phase of the field independently for any point  $x$

$$\phi(x) \rightarrow \phi'(x) = e^{i\Lambda(x)} \phi(x) \quad (2.28)$$

The potential  $V(|\phi|^2) = V(|\phi'|^2)$  is invariant under this transformation but the kinetic term is not

$$\partial_\mu \phi^\dagger \partial_\mu \phi \rightarrow |(\partial_\mu + i(\partial_\mu \Lambda)\phi|^2 \quad (2.29)$$

To compensate this non-invariant change one can introduce new vector field  $A_\mu(x)$  with the transformation law :

$$A_\mu(x) \rightarrow A'_\mu(x) = A_\mu(x) + \frac{1}{e} \partial_\mu \Lambda(x) \quad (2.30)$$

so that the new Lagrangian

$$\mathcal{L} = |(\partial_\mu - ieA_\mu)\phi|^2 - V(|\phi|^2) \quad (2.31)$$

is locally  $U(1)$  invariant or gauge invariant. The combination  $\mathcal{D}_\mu = \partial_\mu - ieA_\mu$  has a name of covariant derivative (or long derivative). It has a simple transformations law

$$\begin{aligned} D_\mu &\rightarrow e^{i\Lambda} \mathcal{D}_\mu e^{-i\Lambda} \\ D_\mu \phi &\rightarrow e^{i\Lambda} (\mathcal{D}_\mu \phi) \end{aligned} \quad (2.32)$$

Up to now the fields  $A_\mu(x)$  have no kinetic term in the Lagrangian and they are some kind of the auxiliary fields that do not propagate. To construct kinetic term we need gauge invariant combination of the derivatives of field  $A_\mu$ . Notice that covariant derivatives and any combinations of the covariant derivatives have a very simple transformation law eq. (2.32). Consider the commutator of two derivatives,

$$\begin{aligned} [\mathcal{D}_\mu \mathcal{D}_\nu] &\equiv -ieF_{\mu\nu} \\ F_{\mu\nu} &= \partial_\mu A_\nu - \partial_\nu A_\mu \end{aligned} \quad (2.33)$$

We see that commutator is not the differential operator but the function of  $x$ . According to (2.32) it is gauge invariant function. Now we are in position to write the total gauge invariant Lagrangian

$$\begin{aligned} \mathcal{L} &= -\frac{1}{4}F_{\mu\nu}F_{\mu\nu} + |\mathcal{D}_\mu \phi|^2 - V(|\phi|^2) \\ \phi(x) &\rightarrow \phi'(x) = e^{i\Lambda(x)} \phi(x) \\ A_\mu(x) &\rightarrow A'_\mu(x) = A_\mu(x) + \frac{1}{e}\partial_\mu \Lambda(x) \end{aligned} \quad (2.34)$$

The notion of gauge invariance was introduced by V. Fock in 1926 and in two steps by H. Weyl in 1919 and 1929.

## 2.4 Spontaneous Breaking of local symmetry : Higgs Phenomenon.

For the case when time derivative is zero  $D_0\phi = 0$  and electric field is zero  $F_{0i} = E_i = 0$  the Lagrangian (2.34) formally is equal to the free energy in the Ginzburg-Landau phenomenological theory of superconductivity, where  $\phi(x)$  plays a role of the order parameter. It is known that magnetic field does not penetrate into superconductor, it falls exponentially. Exponential fall in QFT corresponds to a massive particle. So one can expect that Lagrangian (2.34) at certain circumstances can describe the massive gauge field. This is the famous Higgs mechanism of spontaneous breaking of local symmetry.

Consider Lagrangian (2.34) with the special choice of potential energy (2.32)

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}^2 + |\mathcal{D}_\mu \phi|^2 - \lambda(|\phi|^2 - \frac{\eta^2}{2})^2 \quad (2.35)$$

Potential  $V(\phi)$  has continuous valley of minima. Let us quantize the fields near the vacuum state (2.22)

$$\langle \phi \rangle = \phi_{cl} = \frac{1}{\sqrt{2}}\eta \quad (2.36)$$

As in the case of global symmetry it is convenient to use representation of  $\phi(x)$  in term of modulus and phase

$$\phi(x) = \frac{1}{\sqrt{2}}(\eta + \rho(x))e^{i\sigma(x)} \quad (2.37)$$

The Lagrangians (2.34) and (2.35) are gauge invariant. Then let us make gauge transformation with  $\Lambda(x) \equiv -\sigma(x)$

$$\begin{aligned}\phi(x) &\rightarrow \phi' = e^{i\sigma}\phi \\ A_\mu(x) &\rightarrow A'_\mu = A_\mu - \frac{1}{e}\partial_\mu\sigma\end{aligned}\quad (2.38)$$

In this gauge (unitary gauge)

$$D_\mu\phi \rightarrow (\partial_\mu - ieA_\mu)\frac{1}{\sqrt{2}}(\eta + \rho(x)) \quad (2.39)$$

and the Lagrangian can be rewritten in the form

$$\begin{aligned}\mathcal{L} = & \left[ -\frac{1}{4}F_{\mu\nu}^2 + \frac{1}{2}e^2\eta^2 A_\mu^2 \right] + \\ & \frac{1}{2}(\partial_\mu\rho)^2 + (e^2\eta)\rho(x)A_\mu^2(x) + \frac{e^2}{2}A_\mu^2(x)\rho^2(x)\end{aligned}\quad (2.40)$$

The term in bracket represents the free massive vector particle with mass

$$m_V = e\eta \quad (2.41)$$

Massless Goldstone mode  $\sigma(x)$  has been eaten by massless vector field  $A_\mu(x)$  (that had two polarization) and as a result we get massive vector field with three polarization. This is **Higgs Phenomenon**.

## 2.5 Local $SU(2)$ . Yang-Mills theory of vector fields.

To be ready for the construction of the Standard Model we have to consider the general case of the local gauge groups.

Let us start with  $SU(2)$  theory of massless fermions  $\psi = \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix}$

$$\mathcal{L} = \bar{\psi}[i\gamma_\mu\partial_\mu\psi] \quad (2.42)$$

and consider local  $SU(2)$  transformations

$$\psi(x) \rightarrow \psi'(x) = S(x)\psi(x) \quad (2.43)$$

where

$$\begin{aligned}S(x) &= \exp i(T_j\Lambda_j(x)) ; \\ T_i &= \frac{1}{2}\tau_i \quad , \quad i = 1, 2, 3 ; \\ [T_i, T_j] &= ie_{ijk}T_k\end{aligned}\quad (2.44)$$

The Lagrangian (2.42) is not invariant under this transformation. To compensate the non-invariant piece in the Lagrangian we introduce the triplet of vector fields  $A_\mu^i(x)$  so that:

$$\begin{aligned}\mathcal{L} &= \bar{\psi}i\gamma_\mu(\partial_\mu - igA_\mu(x))\psi \\ A_\mu(x) &= T^i A_\mu^i(x)\end{aligned}\quad (2.45)$$

with the transformation law

$$A_\mu(x) \rightarrow A'_\mu(x) = SA_\mu(x)S^+ - \frac{i}{g}(\partial_\mu S)S^+ \quad (2.46)$$

One can introduce the covariant derivative

$$\mathcal{D}_\mu = \partial_\mu - igA_\mu \quad (2.47)$$

that transforms as a triplet under  $SU(2)$  transformations:

$$\begin{aligned}D_\mu &\rightarrow SD_\mu S^+ \\ D_\mu\psi &\rightarrow S(D_\mu\psi)\end{aligned}\quad (2.48)$$

We can define the triplet of field-strength tensor  $G_{\mu\nu}^i$  :

$$\begin{aligned}
G_{\mu\nu} &\equiv G_{\mu\nu}^i T^i = \frac{i}{g} [\mathcal{D}_\mu, \mathcal{D}_\nu] \\
&= \partial_\mu A_\nu - \partial_\nu A_\mu - ig[A_\mu, A_\nu] \\
G_{\mu\nu} &\rightarrow G'_{\mu\nu} = SG_{\mu\nu}S^\dagger
\end{aligned} \tag{2.49}$$

and construct the  $SU(2)$  gauge invariant Lagrangian

$$\mathcal{L} = -\frac{1}{4} \text{Tr}[G_{\mu\nu} G_{\mu\nu}] + \bar{\psi} i \gamma_\mu \mathcal{D}_\mu \psi \tag{2.50}$$

This Lagrangian was written first time by Yang and Mills in 1954. The very nontrivial part in this construction is that kinetic energy  $\sim G_{\mu\nu}^2$  contains bilinear  $\sim A^2$ , trilinear  $\sim A^3$  and quadrilinear  $\sim A^4$  terms:



Thus Yang-Mills gauge theory is a theory of self-interacting vector fields.

## 2.6 Spontaneous Breaking of Local $SU(2)$ Symmetry : Renormalizable theory of massive vector fields.

Consider the  $SU(2)$  gauge theory of the couple of scalar fields  $\phi = \begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix}$ :

$$\mathcal{L} = -\frac{1}{4} \text{Tr} G_{\mu\nu} G_{\mu\nu} + |\mathcal{D}_\mu \phi|^2 - \lambda(|\phi|^2 - \frac{\eta^2}{2})^2 \tag{2.51}$$

We expect that after spontaneous breaking of  $SU(2)$  symmetry three Goldstone bosons will be mixed with three massless vector fields and produce three massive vector fields.

Let us introduce a special representation for the doublet  $\phi$

$$\phi(x) = \begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix} = e^{i\sigma^a(x)T^a} \begin{pmatrix} 0 \\ \frac{1}{\sqrt{2}}(\eta + \rho(x)) \end{pmatrix} \tag{2.52}$$

and consider gauge transformation with the parameter

$$\Lambda^a(x) = -\sigma^a(x) \tag{2.53}$$

In this gauge the fields  $\sigma^i(x)$  disappear from the Lagrangian and vector part of  $\mathcal{L}$  gets the form

$$\begin{cases} \mathcal{L}_{\text{vect}} = -\frac{1}{4} \text{Tr} G_{\mu\nu}^2 - \frac{1}{2} m_V^2 A_\mu^2 \\ m_V = \frac{1}{2} g \eta \end{cases} \tag{2.54}$$

This is the theory of massive vector fields with the special choice of self-interactions.

The theory of massless Yang-Mills field was renormalizable theory. On the other hand the properties of the vacuum should not change the behavior of the amplitudes at high energy. Thus one can believe that Yang-Mills theory with spontaneous breaking of gauge symmetry remains renormalizable. The theory of massive vector fields with arbitrary interactions is non-renormalizable in general. But if one takes the special case of interaction with quarks, with scalars and self-interaction that corresponds to the gauge-invariant Lagrangian (2.51) the non-renormalizable divergences should disappear. Technically the rigorous proof of this statement is quite nontrivial business even now. This problem had been solved by t'Hooft and Veltman in 1971.

### 3 $SU(2) \times U(1)$ Theory of Electroweak Interactions.

There is no unique way to construct the theory of electroweak interactions. In 1970s there were several dozens of models on the market. Only the simplest one has survived in our time. A priori we do not understand why the gauge group is  $SU(2) \times U(1)$ , why there are three generations of quarks and leptons etc. All these questions have no answer inside SM. We have to look for the answers into experiments.

It was well established in the old four-fermion theory of weak interactions that charged currents (responsible for  $\beta$ -decay of leptons and hadrons) have  $V - A$  structure, i.e. they are constructed from the left-handed fermions.

The minimal group of gauge symmetry which includes charged vector currents is  $SU(2)$  group. Thus any theory of weak interactions have to include  $SU(2)_L$  symmetry as a subgroup. Photon interacts both with left- and right-handed fermions. To unify weak and electromagnetic interactions the group of gauge symmetry should include  $U(1)$  as well. The minimal group of symmetry that includes these subgroups is the product

$$G = SU(2)_L \times U(1).$$

#### 3.1 Left and Right Fermions. $SU(2)_L$ symmetry.

Any Dirac 4-spinor  $\Psi$  can be presented as a sum of two Weyl spinors  $\Psi_L$  and  $\Psi_R$ :

$$\Psi = \Psi_L \oplus \Psi_R = \frac{1}{2}(1 + \gamma_5)\Psi + \frac{1}{2}(1 - \gamma_5)\Psi \quad (3.1)$$

Two-component Weyl spinors are irreducible representations of Lorentz group. For massless fermions

$$\Psi_L = \begin{pmatrix} (1 - \vec{\sigma}\vec{n})\varphi \\ -(1 + \vec{\sigma}\vec{n})\varphi \end{pmatrix}, \quad (3.2)$$

where  $\varphi$  is 2-spinor,  $\vec{\sigma}$  are Pauli matrices, and  $\vec{n} = \vec{p}/|p|$  is the direction of the motion of particle. So for left particle  $\Psi_L$

$$\vec{\sigma}\vec{n} = -1 \quad (3.3)$$

and for right particles  $\Psi_R$

$$\vec{\sigma}\vec{n} = +1 \quad (3.4)$$

Left leptons and quarks group into  $SU(2)_L$  doublets. For the first generations they are

$$L = \begin{pmatrix} \nu_e \\ e \end{pmatrix}_L \quad \text{and} \quad Q = \begin{pmatrix} u \\ d \end{pmatrix}_L \quad (3.5)$$

To avoid  $V + A$  charged current we have to put right fermions into singlet representation. Thus  $e_R$ ,  $u_R$  and  $d_R$  are singlets. As for right-handed neutrino  $\nu_R$  nobody has observed it so far. It is unknown whether such field exists. Just now we prefer not to introduce  $\nu_R$  into the theory.

To include the electromagnetic interactions we have to define charge. For left-handed fermions the charge is different for up and down component. Thus

$$Q_L = T_3 + Y_L \quad (3.6)$$

where  $T_3$  is the third component of  $SU(2)_L$  and  $Y_L$  is left hypercharge (for leptonic doublet  $Y_L = -1/2$  and for quark doublet  $Y_Q = -1/6$ ).

For right fermions we identify  $Y_R$  and  $Q$ :

$$Q_R = Y_R \quad (3.7)$$

so that  $Y_{u_R} = \frac{2}{3}$ ;  $Y_{d_R} = -\frac{1}{3}$ ,  $Y_{e_R} = -1$ .

The minimal way to introduce  $U(1)$  interactions is to consider gauge boson that interacts with hypercharge  $Y$

$$Y = Y_L + Y_R \quad (3.8)$$

This is the gauge group of Minimal Standard Model

$$SU(2)_L \times U(1)_Y \quad (3.9)$$

Let  $A_\mu^i(x)$ ,  $i = 1, 2, 3$  be gauge bosons of  $SU(2)_L$  and  $B_\mu(x)$  – the gauge boson of  $U(1)$  group. The charged fields

$$A_\mu^\pm = \frac{1}{\sqrt{2}}(A_\mu^1 \pm iA_\mu^2) \quad (3.10)$$

can be identify with  $W_\mu^\pm$  bosons.

Photon  $A_\mu(x)$  in general is a combination of  $A_\mu^3$  and  $B_\mu$ . Orthogonal combination represents another physical state, i.e.  $Z$  boson. Thus

$$\begin{pmatrix} Z_\mu \\ A_\mu \end{pmatrix} = \begin{pmatrix} \cos \theta_W & -\sin \theta_W \\ \sin \theta_W & \cos \theta_W \end{pmatrix} \begin{pmatrix} A_\mu^3 \\ B_\mu \end{pmatrix} \quad (3.11)$$

where  $\theta_W$  is a weak mixing angle.

To break spontaneously  $SU(2)_L \times U(1)_Y$  group and to make masses to  $W^\pm$  and  $Z$  bosons we need three Goldstone fields. The  $SU(2)$  doublet of Higgs particles

$$H = \begin{pmatrix} H^+ \\ H^0 \end{pmatrix}; \quad Y_H = \frac{1}{2} \quad (3.12)$$

can provide three Goldstone bosons after spontaneous breaking. In the MSM we use only **one** Higgs doublet.

We have completed the construction of the MSM. Now we are ready to determine the masses of vector bosons  $m_W$ ,  $m_Z$  and phenomenological mixing angle  $\theta_W$  in terms of  $SU(2)$  coupling constant -  $g_2$ ,  $U(1)$  coupling constant -  $g_1$  and in terms of v.e.v. of Higgs field -  $\eta$ .

In the unitary gauge Higgs doublet has the form

$$H(x) = e^{i\vec{T}\vec{a}(x)} \begin{pmatrix} 0 \\ \frac{1}{\sqrt{2}}(\eta + \rho(x)) \end{pmatrix} \quad (3.13)$$

Covariant derivative

$$D_\mu \equiv \partial_\mu - ig_1 Y B_\mu(x) - ig_2 T^a A_\mu^a(x) \quad (3.14)$$

for the vacuum field  $H_{vac}$  is equal to

$$D_\mu H_{vac} = (-ig_1 \frac{1}{2} B_\mu - ig_2 \frac{1}{2} \tau^a A_\mu^a) \begin{pmatrix} 0 \\ \frac{\eta}{\sqrt{2}} \end{pmatrix} = \quad (3.15)$$

$$= \frac{(-i)}{2\sqrt{2}} \eta \begin{pmatrix} \sqrt{2}g_2 W_\mu^- \\ -g_2 A_\mu^3 + g_1 B_\mu \end{pmatrix}$$

The mass term for vector fields originates from  $(D_\mu H)^\dagger D_\mu H$  term in the Lagrangian. It looks like

$$\mathcal{L}_{mass} = \frac{1}{4}(g_2 \eta)^2 W_\mu^+ W_\mu^- + \frac{1}{8} \eta^2 (g_2 A_\mu^3 - g_1 B_\mu)^2 \quad (3.16)$$

From this expression we conclude that the massive combination of  $A_\mu^3$  and  $B_\mu$  (i.e.  $Z$ -boson) is

$$Z_\mu = \frac{1}{\sqrt{g_1^2 + g_2^2}}(g_2 A_\mu^3 - g_1 B_\mu) \quad (3.17)$$



or that

$$tg\theta_W = g_1/g_2 \quad (3.18)$$

From eq. (3.16) it follows that

$$m_W = \frac{1}{2}g_2\eta \quad (3.19)$$

and

$$m_W = m_Z \cos \theta_W \quad (3.20)$$

It is very interesting that  $Z$  boson should be heavier than  $W$  boson! After spontaneous breaking there still remains unbroken  $U(1)$  symmetry that corresponds to massless photon.

If we introduce electric charge  $e$  as a coupling constant of the photon we can relate  $g_{1,2}$  with  $e$  and  $\cos \theta_W$ . Let us rewrite interaction of  $A_\mu^3$  and  $B_\mu$  as an interaction of  $A_\mu$  and  $Z_\mu$  fields:

$$(-ig_2T_3)A_\mu^3 - ig_1YB_\mu \equiv (-i)\frac{g_2}{\cos \theta_W}[T_3 - \sin^2 \theta_W Q]Z_\mu + (-i)(g_1 \cos \theta_W)QA_\mu \quad (3.21)$$

This is identically rewritten universal expression for covariant derivative. So eq. (3.21) is applicable to the left and right fermions and to the Higgs doublet.

From eq. (3.21) it follows immediately that

$$e = g_1 \cos \theta_W = g_2 \sin \theta_W \quad (3.22)$$

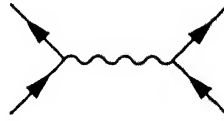
We complete the description of bosonic sector of the SM.

### 3.2 Weak interactions of leptons and quarks : Neutral Current. Request for new particles.

Now we are ready to calculate the amplitude for the first physical process, for the decay of  $\mu \rightarrow e\nu\bar{\nu}$ . Charged currents Lagrangian for leptons looks like

$$\Delta\mathcal{L}_{Charged} = \frac{g_2}{2\sqrt{2}}W_\mu^+[\bar{\nu}\gamma_\mu(1+\gamma_5)e + \dots] \quad (3.23)$$

where the dots are for the similar terms with  $\mu$  and  $\tau$  leptons. Feynman diagram for the  $\mu$ -decay is



The amplitude for the decay can be read from this diagram and it is equal to

$$T(\mu \rightarrow e\nu\bar{\nu}) = [\frac{g_2}{2\sqrt{2}}]^2 \frac{1}{m_W^2 - q^2} (\bar{\nu}\gamma_\alpha(1+\gamma_5)\mu)(\bar{e}\gamma_\alpha(1+\gamma_5)\nu) \quad (3.24)$$

The momentum transfer  $q$  from muonic current to electronic current is of the order of muonic mass  $m_\mu$ . So if  $m_W \gg m_\mu$  the amplitude looks like a point-like interaction in Fermi theory.

$$T_{Fermi} = \frac{G_F}{\sqrt{2}}j_\alpha^e(j_\alpha^\mu)^+ \quad (3.25)$$

Comparing these two presentations for the same amplitude we conclude that

$$\frac{G_F}{\sqrt{2}} = \frac{g_2^2}{8m_W^2} \quad (3.26)$$

Taking into account eq. (3.19) for  $m_W$  we also get that v.e.v.  $\eta$  is directly connected with  $G_F$ :

$$\eta = [\sqrt{2}G_F]^{-1/2} = 246 \text{ GeV} \quad (3.27)$$

$$G_F \equiv G_\mu = 1.16639(2) \cdot 10^{-5} \text{ GeV}^{-2}$$

To fix remaining two fundamental parameters  $g_1$  and  $g_2$  we have to choose two other physical observables measured with the best accuracy. The choice is evident. They are the fine coupling constant  $\alpha$

$$\alpha^{-1} = \frac{4\pi}{e^2} = 137.035985(61) \quad (3.28)$$

and the mass of  $Z$ -boson

$$m_Z = 91.187(2) \text{ GeV} \quad (3.29)$$

To calculate  $g_1$  and  $g_2$  we first have to calculate the mixing angle  $\theta_W$  in terms of  $G_F$ ,  $\alpha$  and  $m_Z$ . It is not difficult exercise to show that

$$\sin^2 \theta_W \cos^2 \theta_W = \frac{\pi\alpha}{\sqrt{2}(G_F m_Z^2)} \quad (3.30)$$

Substituting the values of the parameters from eqs. (3.27), (3.28) and (3.30) we get

$$\begin{aligned} \sin^2 \theta_W &= 0.2120 \\ g_1 &= \frac{\sqrt{4\pi\alpha}}{\cos \theta_W} = 0.34 \\ g_2 &= \frac{\sqrt{4\pi\alpha}}{\sin \theta} = 0.66 \end{aligned} \quad (3.31)$$

So we are ready for the first prediction in SM: we can calculate  $m_W$

$$(m_W)^{theor} = m_Z \cos \theta_W = 80.94 \text{ GeV} \quad (3.32)$$

that has to be compared with the current experimental value

$$(m_W)^{exp} = 80.37(8) \text{ GeV} \quad (3.33)$$

The deviation from theoretical number is only 0.6%, but this tiny number is equal to  $8\sigma$  deviation. To explain the huge discrepancy we have to take into account radiative correction that have the scale of the few per mill.

The old 4-fermionic point-like theory is the effective theory for momentum transfer much smaller than  $m_W$ . In this sense the SM is generalization of the old theory. But SM also predicts the new phenomena that were unknown in V-A theory. This is the neutral currents.

The effective 4-fermionic coupling of neutral currents is generated by  $Z$  boson exchange.



At small momentum transfer it is local interaction with the coupling constant equal to  $G_F \cos^2 \theta_W$ .

Though this coupling is of the same order as  $G_F$  by some reasons the experimental search for neutral currents gave negative results for a long time and only in 1973 experimental groups at CERN observed neutral currents and provided the first experimental measurements of  $\cos \theta_W$ . This measurement gave the possibility to calculate  $m_W$  and  $m_Z$  theoretically (eqs. (3.19), (3.20)) with

rather good accuracy. This estimate had been extremely helpful for the experimental discovery of  $W$  and  $Z$  bosons.

Another great achievement of the SM was the request for new particles needed for self-consistency of the theory. In 1970 the set of the known particle included 2 generations of leptons

$$\begin{pmatrix} \nu_e \\ e \end{pmatrix}_L, \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix}_L; e_R, \mu_R \quad (3.34)$$

and three quarks  $u$ ,  $d$  and  $s$  that belong to the following  $SU(2)_L \times U(1)_Y$  representation

$$\begin{pmatrix} u \\ d' = d \cos \theta_c + s \sin \theta_c \end{pmatrix}_L, u_R, d_R, s_R \quad (3.35)$$

where  $\theta_c$  is the Cabibbo angle. This set of quarks produces flavor-changing  $s \leftrightarrow d$  neutral currents

$$\begin{aligned} Z_\mu \bar{d}'_L \gamma_\mu d'_L &\sim Z_\mu [(\bar{d}d) \cos^2 \theta_c + (\bar{s}s) \sin^2 \theta_c + \\ &+ \sin \theta_c \cos \theta_c (\bar{d}s + \bar{s}d)] \end{aligned} \quad (3.36)$$

This was absolutely forbidden by experimental data. To save the SM Glashow, Illiopoulos and Maiani in 1970 introduced fourth  $c$  quark and the new  $SU(2)_L$  doublet

$$\begin{pmatrix} c \\ -d \sin \theta_c + s \cos \theta_c \end{pmatrix} \quad (3.37)$$

As a result flavor-changing neutral currents disappear and all neutral currents become diagonal. This theoretical request for new particle was satisfied by experimental discovery of  $c$ -quark in 1974. The second generation had been completed.

The  $\tau$ -lepton and beauty  $b$ -quark were discovered experimentally in mid-1970s. The heaviest particle - top  $t$ -quark was discovered only two decades later. They compose the third generation of matter. Third generation is absolutely necessary to describe the experimental data with high accuracy, i.e. in loop approximation. Moreover to fit experimental data, the top quark mass should be equal to

$$m_t = 180(5)_{-20}^{+17} \text{ GeV}$$

That is very close to the actual experimental value of top mass.

### 3.3 Quark masses.

In the Standard Model the mass term for the electron violates  $SU(2)_L$ . Indeed this term

$$m_e \bar{e}e = m_e [\bar{e}_R e_L + \bar{e}_L e_R] \quad (3.38)$$

transforms like doublet instead of being invariant.

To preserve  $SU(2)_L \times U(1)_Y$  symmetry we have to use Higgs mechanism to generate the masses for fermions. For example Yukawa coupling of  $L$ ,  $e_R$  and  $H$  is  $SU(2)_L \times U(1)_Y$  invariant

$$\begin{aligned} \Delta \mathcal{L} &= f_e (\bar{L} e_R) H + h.c. = \\ &= \frac{f_e}{\sqrt{2}} (\eta + \rho(x)) \bar{e}e = \\ &= m_e \bar{e}e + \frac{f_e}{\sqrt{2}} \rho(x) \bar{e}e \end{aligned} \quad (3.39)$$

where  $\rho(x)$  is the field for physical Higgs in SM. From eq. (3.39) it follows that Yukawa coupling is proportional to  $m_e$

$$f_e = \frac{\sqrt{2}}{\eta} m_e \simeq 3 \cdot 10^{-6} \quad (3.40)$$

Notice that before this step the fields  $e_L(x)$  and  $e_R(x)$  were absolutely different, i.e. they had different interaction with  $W$  and  $Z$ . Yukawa interaction unified these two Weyl spinors into one massive particle – electron. To give the mass to down quarks we can use the same type of Yukawa interaction

$$\Delta\mathcal{L}_m = f_d(\bar{Q}_L d_R)H \quad (3.41)$$

As for the mass of up quarks we need Higgs doublet with nonzero v.e.v. for up component of doublet. At that moment we can introduce new Higgs doublet. But in the case of  $SU(2)$  group complex conjugated fields

$$\tilde{H} = (-i\sigma_2)H^* \quad (3.42)$$

also behave like a member of  $SU(2)$  doublet. So we can use  $\tilde{H}$  to give mass to upper quark

$$\Delta\mathcal{L}_m = f_d(\bar{Q}_L d_R)H + f_u(\bar{Q}_L u_R)\tilde{H} \quad (3.43)$$

This is the solution of problem of fermion mass in the case of one generation.

For more than one generation we have to take into account quark mixing. For two generations this mixing can be described by one rotation angle (Cabibbo angle). For three generations one gets three independent rotations and one complex phase. Complex couplings cause violation of CP-invariance. Thus in the theory with tree generations we get mechanism for CP-breaking. Till now it remains unknown whether this mechanism is the only source of CP-violation in Nature.

### 3.4 Triangle Anomaly.

To have renormalizable theory of electroweak interactions it was absolutely crucial to start from the gauge invariant theory where gauge bosons interact with conserved Noether currents. Spontaneous breaking of symmetry does not spoil any symmetric relations between operators. They are exactly the same as in the original theory. The confusing notion of spontaneous breaking describes the nonlinear realization of the symmetry in the space of physical states.

In the SM we operate both with vector and axial currents. For any axial currents

$$j_\mu^A = \bar{\Psi}\gamma_\mu\gamma_5\Psi \quad (3.44)$$

$$\partial_\mu j_\mu^A = 2im\bar{\Psi}\gamma_5\Psi$$

Thus naively axial current is conserved for massless fermions. But what is true in Classical Field Theory can be not true in Quantum Field Theory. Indeed one-loop calculation of the divergence of axial current for electrons gives

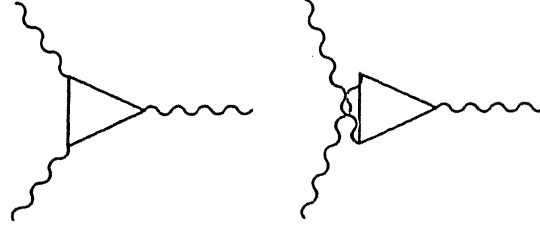
$$\begin{cases} \partial_\mu j_\mu^5 = 2im\bar{\Psi}_e\gamma_5\Psi_e + \frac{\alpha}{2\pi}F_{\mu\nu}\tilde{F}_{\mu\nu} \\ \tilde{F}_{\mu\nu} = \frac{1}{2}\varepsilon_{\mu\nu\alpha\beta}F_{\alpha\beta} \end{cases} \quad (3.45)$$

instead of eq. (3.44). The term  $F\tilde{F}$  originates from matrix element of  $\partial_\mu j_\mu^5$  between vacuum and two-photon states.

So the axial current is not conserved even for  $m \equiv 0$ . Not any classical symmetry can survive in Quantum Mechanics. This very interesting theoretical phenomenon has special name – triangle anomaly.

In the SM there are plenty of axial currents that interact with gauge fields. Though fermions are massless (no mass terms in the Lagrangian) the anomaly can destroy the conservation of Noether currents and this will kill renormalizability. There is one possibility to save it. We see from eq. (3.45) that anomaly depends only on the "charge" of particle that is running inside loop. So if the total gauge current has different pieces it can happen that nonzero individual anomalies cancel each other for the total current.

This cancellation imposes some restrictions on the charges of quarks and leptons. Let us check this possibility. We will calculate the triangle matrix elements between fields  $A_\mu^i$  and  $B_\mu$ . There are two crossing diagrams that contribute to anomalous interaction between 3 gauge fields.



Consider first the anomalous contribution of one generation of matter. Rather simple calculations demonstrate the following result.

- 1)  $(A, A, A)$  and  $(A, B, B)$  anomalies are automatically disappeared for lepton doublet and for quark doublet separately.
- 2)  $(B, A, A)$  anomaly is disappeared if

$$Q_e + 2Q_u + Q_d \equiv 0, \quad (3.46)$$

i.e. quark contribution cancels lepton contribution only for this special relation between charges. This relation means that hydrogen atom has to be neutral!

It is very interesting that renormalizability of the SM takes place only if the charge of proton is opposite to the charge of electron.

We can proceed further and consider other anomalies. At that moment we have to make some statement about  $\nu_R$ . Suppose first that it does not exist at all. In this case:

- 3) Cancellation of  $(B, B, B)$  anomaly takes place only if

$$Q_e = -1, \quad Q_\nu = 0; \quad Q_u = \frac{2}{3}, \quad Q_d = -\frac{1}{3} \quad (3.47)$$

(We suppose that QCD has  $SU(3)_c$  symmetry.)

- 4) Cancellation of  $(B \rightarrow \text{gluon} + \text{gluon})$  anomaly is automatic.

- 5) Cancellation of  $(B \rightarrow \text{graviton} + \text{graviton})$  anomaly takes place only for the charge sample eq. (3.47). So we are able to fix the relative charges of leptons and quarks in this case.

If  $\nu_R$  does exist anomalies 3) - 5) are disappeared automatically for any charge of neutrino.

If we suppose that the new generations are the exact replica of the old one (only masses are different, but the charges are the same) then we come to the same conclusion for each generation. If we allow to change the charges from one generation to another one the restrictions on the quark and lepton charges become weaker.

In any case it is very interesting that renormalizability imposes restrictions on the property of matter fields.

## 4 Loops in SM - the Window into New Physics.

### 4.1 Z-physics at LEP and SLC.

To test the predictions of the SM the huge "factories" of  $Z$ -bosons ( $e^+e^-$  colliders) were constructed at CERN (LEP) and at SLAC (SLC). Electrons and positrons in thus colliders collide at the center of mass energy equal to the  $Z$ -boson mass. The reactions that are studied can be presented in the form

$$e^+e^- \rightarrow Z \rightarrow \bar{f}f$$

where

$$\bar{f}f = \begin{cases} \nu\bar{\nu} & \text{invisible modes} \\ l\bar{l} & \text{charged leptons} \\ q\bar{q} & \text{hadrons.} \end{cases}$$

Near the dozen of independent observables were measured with fantastic precision of the order of  $10^{-3}$  ( $10^{-5}$  for the case of  $Z$ -boson mass). The scale of the radiative corrections in the SM is of the order of weak coupling constants:  $\alpha_{W,Z}/\pi \sim 10^{-2} - 10^{-3}$ . Therefore LEP-I and SLD data provide precision test of the SM as a renormalizable field theory, i.e. with loops included.

The theoretical study of electroweak corrections in SM started in 1970's and was elaborated by a number of theoretical groups. The deviations in theoretical calculations of different groups are by the order of magnitude smaller than the experimental uncertainties.

By comparing the  $\Gamma_{invisible}$  with theoretical predictions for neutrino decays the result of fundamental importance was established – the sequence of the generations with light neutrino is completed with number of generation

$$N_f = 3.$$

The fit of the most recent experimental data (summer 1999) has very good quality:

$$\chi^2/n.d.f. = 15.0/14.$$

We conclude that the SM gives the perfect description of  $Z$  physics. New physics can not improve the fit of LEP and SLC data. Thus the Standard Model has been confirmed up to the loop corrections.

What is more important is that the loop corrections can be used to gather data on the not yet discovered particle. For instance, even before  $t$ -quark was discovered at Tevatron, its mass was predicted by analyzing the loops and LEP-SLC data. The hunting for virtual top quark is a very bright example of the collaboration of the theory and the experiment.

## 4.2 Decoupling of heavy flavors from Low-Energy : Physics in QED and QCD.

It is interesting to understand why in 1950's nobody worried about the contribution of top quark (and other heavy flavors) into magnetic moment of the electron known with very high accuracy. The answer to this question is that for momenta  $q \sim m_e$  the corrections due to top quark are suppressed as a power of  $(m_e^2/m_t^2)$  i.e. the contribution was negligible. In QED any heavy particles decouple from the low-energy observables.

Consider the contribution of  $t$ -quark into QED observables. The only diagram with  $t$ -quarks in loop is the self-energy of the photon

$$\Pi_{\mu\nu}(q) = i\langle 0 | \{j_\mu(q), j_\nu(-q)\} | 0 \rangle \quad (4.1)$$

where  $j_\mu(q)$  is the electromagnetic current of  $t$ -quark. Self energy has dimension 2:  $[\Pi_{\mu\nu}] = m^2$ . So one can expect that there exit terms of the order of

$$\Pi_{\mu\nu} \sim \alpha m_t^2 g_{\mu\nu}$$

This expectation is wrong in the case of conserved currents

$$q_\mu j_\mu(q) = 0 \quad (4.2)$$

Indeed for conserved current the self-energy operator should be transversal  $q_\mu \Pi_{\mu\nu} = 0$ . Thus

$$\Pi_{\mu\nu}(q) = (q_\mu q_\nu q^2 - q_\mu q_\nu) \Pi(q^2), \quad (4.3)$$

Equation (4.3) implies that the photon remains massless. The scalar function  $\Pi(q^2)$  has dimension zero and the only possible contribution of  $t$ -quark into  $\Pi(q^2)$  can be written in the form

$$\Pi(q^2) \sim \alpha \ln \frac{\Lambda^2}{m_t^2 + q^2}$$

where  $\Lambda$  is cut-off. The self-energy keeps the memory of heavy flavors!

The crucial step is renormalization. Consider the example of Coulomb scattering from Chapter I more carefully. If we take into account the infinite chain of self-energy contribution into the photon propagator we get for amplitude

$$T_{Coulomb} = \frac{e_0^2(\Lambda)}{q^2(1 + \Pi(q^2))} \quad (4.4)$$

At low  $q^2$  we reproduce the Coulomb-law

$$T = \frac{e_{phys}^2}{q^2} \quad (4.5)$$

with

$$e_{phys}^2 = \frac{e_0^2(\Lambda)}{1 + \Pi(0)} \quad (4.6)$$

When we rewrite the amplitude (4.4) in terms of  $e_{phys}^2$  we get

$$T \simeq \frac{e_{phys}^2}{q^2[1 + \Pi(q^2) - \Pi(0)]} \quad (4.7)$$

As a result:

1) the dependence on cut-off  $\Lambda$  disappears

$$\Delta\Pi = \Pi(q^2) - \Pi(0) \sim \alpha \ln \frac{m_t^2}{m_t^2 + q^2};$$

2) the contribution of heavy flavor is suppressed as a power ( $q^2/m_t^2$ ):

$$\Delta\Pi \sim -\alpha \left( \frac{q^2}{m_t^2} \right) \rightarrow 0.$$

This is so called **decoupling theorem**. It works for the theories with conserved vector currents.

### 4.3 Non-decoupling of chiral matter :

#### Heavy Flavor contribution into electroweak observables.

In the Standard Model the left components of  $t$ - and  $b$ -quarks belong to  $SU(2)_W$  doublet representation:  $Q_L = \begin{pmatrix} t_L \\ b_L \end{pmatrix}$ . Therefore for the case when  $m_t \gg m_b$  and for small energies  $E \leq m_t$  we have effectively the explicit violation of  $SU(2)_W$  symmetry. For the virtual momenta  $q \sim \Lambda \sim m_t$  theory looks like the old non-renormalizable theory. It mean that one-loop corrections diverge quadratically  $\delta_1 \sim \alpha\Lambda^2/m_Z^2$ , two-loop corrections diverge quartically  $\delta_2 \sim (\alpha\Lambda^2/m_Z^2)$ . So we expect that the corrections to the low-energy observables due to top contribution are of the order of

$$\delta_1 \sim \alpha_W t$$

$$\delta_2 \sim \alpha_W^2 t^2$$

where  $t = m_t^2/m_Z^2$ , i.e. corrections are not suppressed, they grow with top mass  $m_t$ . Heavy flavors are not decoupled from the low-energy observables for the chiral matter. As a result the radiative

corrections in the SM are sensitive to the top contribution. Hunting for virtual top was very successful.

A comparison of the experimental data with the result of theoretical calculation led to the prediction of the  $t$ -quark mass  $m_t$ :

$$m_t = 180(5)_{-20}^{+17} \text{ GeV},$$

where the number in parentheses is the uncertainty due the uncertainties of the data. The center value corresponds to the assumption that  $m_H = 300 \text{ GeV}$ , the upper and lower "shifts" correspond to  $m_H = 1000 \text{ GeV}$  and  $60 \text{ GeV}$ , respectively.

The best fit of all observables gives

$$(m_t) \simeq 170.6 \pm 4.9 \text{ GeV}$$

These numbers are in perfect agreement with the recent direct measurement of the top-quark mass by two collaboration at FNAL

$$m_t = 173.8 \pm 5.0 \text{ GeV}.$$

The same strategy works for heavy unknown particles. Direct accelerator searches did not find yet any trace of New Physics. But loops with hypothetical new particles change the predictions of the SM. Thus the indirect way to get information on New Physics is to study the low-energy processes with very high accuracy.

The simplest extension of the SM is the SM with new sequential generations of heavy quarks and leptons. Chiral matter does not decouple from low-energy observables even when the masses of particles become very large. If the masses of up and down components are very different we get effectively the violation of  $SU(2)$  symmetry and the corrections are large. They are similar to the top quark contribution. If the masses are heavy but degenerate the contribution into observables does not grow with mass but remains finite and nonzero. Accurate analysis demonstrates that one extra generation is already excluded by the current experimental data!

#### 4.4 Hunting for virtual Higgs.

The direct search for Higgs at accelerators still gives negative result. The recent experimental lower bound for Higgs mass is near  $90 \text{ GeV}$ . In this situation it seems reasonable to look for virtual Higgs in loops.

Consider the limit of very large Higgs boson mass  $m_H$ . For  $E \ll m_H$  we have  $SU(2)$  symmetric theory of massive gauge bosons, i.e. effectively non-renormalizable theory. Due to the gauge symmetry the leading divergence of the loop disappears. So the one-loop corrections diverge logarithmically

$$\delta_1 \sim \alpha_W \ln \frac{\Lambda^2}{m_Z^2} \sim \alpha_W \ln \frac{m_H^2}{m_Z^2} = \alpha_W \ln h$$

two-loop corrections diverge quadratically

$$\delta_2 \sim \alpha_W^2 \left( \frac{\Lambda^2}{m_Z^2} \right) \sim \alpha_W^2 h.$$

Here  $h = m_H^2/m_Z^2$ .

This is the famous Veltman screening theorem. The weak dependence of radiative corrections on  $h$  results in a rather poor accuracy for  $m_H$  derived from the precision data. The central value of  $m_H$  from the fit is very unstable. Any tiny corrections or any change of the parameter can shift it by the order of magnitude. The one sigma upper bound is more reliable. According to the recent fit

$$m_H = 70.8_{-43}^{+88} \text{ GeV}.$$

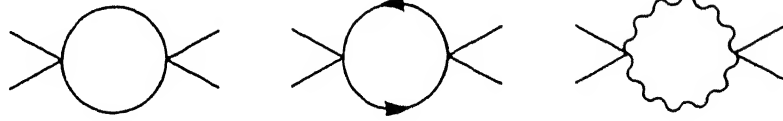
It seems that the fit of the precision data is not the best place to hunt for Higgs boson.



#### 4.5 Effective potential. Stability of the Universe and Bounds on $m_H$ : The Higgs potential in the SM

$$V_{cl}(H) = \frac{\lambda}{4} H^4 - \frac{\mu^2}{2} H^2 \quad (4.25)$$

has minima that corresponds to nonzero v.e.v. of field  $H$ :  $\langle H \rangle_{vac} = \eta$ . Loop corrections change self-interactions of Higgs particles:



The effective potential with loop corrections was calculated by Coleman and Weinberg in 1973. In one-loop approximation it looks like

$$V_{eff}(H) - V_{cl}(H) = \frac{1}{64\pi^2} \left\{ \frac{m_H^4 + 6m_W^4 + 3m_Z^4}{\eta^4} - \frac{12m_t^4}{\eta^4} \right\} H^4 \ln \frac{H^2}{M^2} \quad (4.26)$$

where we have neglected by small contributions from fermions other than  $t$ -quark. Note that due to Fermi-Dirac statistic the contribution of fermion loops has opposite sign in compared to the bosonic loops.

Corrections (4.26) become more important than the main classical potential (4.25) for very large field  $H$ .

In one-loop approximation we get that the correction (4.26) has negative sign if  $m_H < \sqrt[4]{12} m_t$ . For this case the effective potential has no ground state (see Fig.2). Thus even if our system was located first in the local minima at  $\langle H \rangle = \eta$  it will decay at  $t \rightarrow \infty$  and the average value of field  $H$  will run to infinity. We know that nothing like that has happened with our Universe that is near  $10^{10}$  years old. So the stability of the Universe imposes strong constraints on the masses of top and Higgs particles.

To get more reliable results we have to improve a little one-loop formula (eq. (4.26)). For large  $H$  one-loop logarithmic corrections  $\lambda \ln H$  and  $\alpha_W \ln H$  are of the same order as tree terms, two-loop double-logarithmic terms are of the order of one-loop terms etc. So all these logarithmic terms should be taken into account. This technical problem is not very difficult - renormalization group techniques help to sum up such corrections. The result is

$$V^{eff}(H) = -\frac{1}{2} \mu^2(t) H^2(t) + \frac{1}{4} \lambda(t) H^4(t) \quad (4.27)$$

where  $\lambda(t)$  and  $\mu(t)$  are running parameters and  $t = \ln H/\eta$ . For small value of  $t$  (i.e. for small value of field  $H$ ) the running parameters do not run far away from their classical values and the effective potential is equal to the classical one with the accuracy of small radiative corrections. For large  $H$  we can forget about  $\mu^2 H^2$  and the whole dynamics at large  $H$  is governed by running coupling constant  $\lambda(t)$ . There are different contributions into running of  $\lambda(t)$  coming from the loops with top quarks, vector bosons and Higgs boson itself. If the top quark contribution dominates, i.e. the Higgs coupling to top (i.e. the top mass) is large,  $\lambda(t)$  changes sign and the vacuum becomes unstable. This is reformulation of the phenomena that we had at one loop level.

If the Higgs self-interaction dominates, i.e. the Higgs mass is large, then the evolution of  $\lambda(t)$  is similar to the evolution of coupling in the  $H^4$  theory without other fields. It is known that in this case the behavior of  $\lambda(t)$  is

$$\lambda(t) = \frac{\lambda(t_0)}{1 - b\lambda(t_0) \ln \frac{t}{t_0}} \quad (4.28)$$

and running coupling goes to infinity at some finite value of  $H$ .

$$H = \Lambda$$

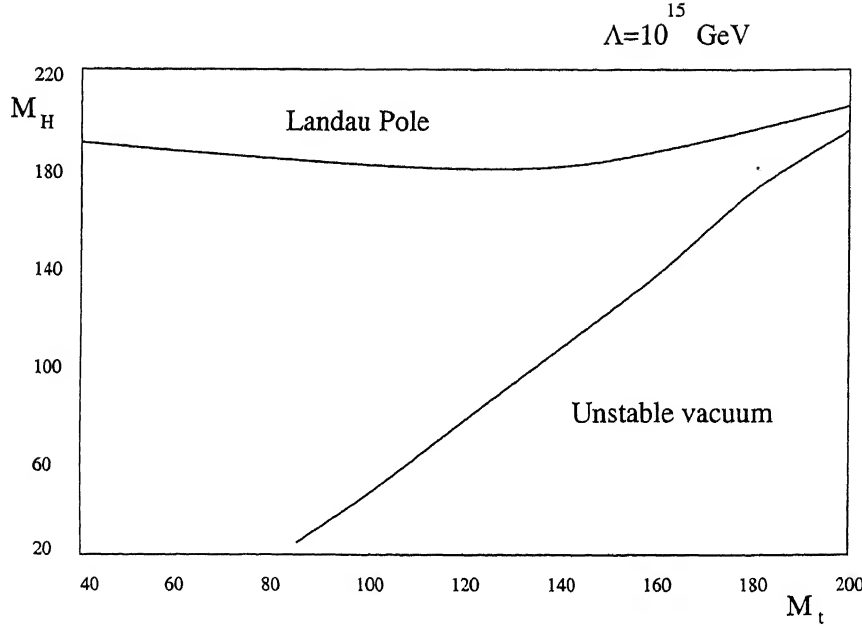


Figure 2:

This is the Landau pole in the running coupling constant. When initial condition  $\lambda(t_0)$  (i.e. the value of Higgs mass) increases the value of Landau pole goes down. If we substitute this running coupling constant into eq.(4.27) we get that the effective potential runs to infinity at this value of  $H$  as well. Such singular behavior of the coupling constant is unacceptable from the physical point of view. Indeed for any finite value of the bare coupling constant  $\lambda^B$  ( $\lambda^B$  is equal to the running coupling  $\lambda(t)$  at the cut-off  $\Lambda$ ) we get that renormalized coupling constant (i.e.  $\lambda(t)$  at low value of  $t$ ) is equal to zero. It means that at low energy we get trivial free theory. This pathological theory seems to be unphysical.

There is possibility to improve that bad behavior of  $\lambda(t)$ . If some new physics (i.e. new interactions and new particle) contribute into  $\lambda(t)$  at scale below or near  $\Lambda$  the pole can disappear.

If we believe that there are no new physics up to some scale (or that the theory can be treated perturbatively up to this scale) we have to push the position of Landau pole  $\Lambda$  (calculated in one-loop approximation eq. (4.28)) to higher scale. This impose upper bound on the value of Higgs mass. So we have bounded  $m_H$  both from above and from below. This remarkable line of reasoning was invented by Cabibbo et al. in 1979.

There are different choices for the parameter  $\Lambda$ . For example  $\Lambda$  can be of the order of Planck scale

$$\Lambda \sim \Lambda_{Pl} = 10^{19} \text{ GeV} ,$$

or of the order of Grand Unification Scale

$$\Lambda \sim \Lambda_{GUT} = 10^{15} \text{ GeV} ,$$

or of the scale of the energy of the accelerator of the next generation

$$\Lambda \sim 10^3 - 10^5 \text{ GeV} .$$

We have to keep in mind all these possibilities. It is evident that for the strongest assumption that new physics does not appear up to the Planck scale we should get the strongest upper bound for  $m_H$ .

To derive more quantitative results we have to solve differential equation for the running coupling constant  $\lambda(t)$ . The renormalization of  $\lambda(t)$  depends on self-interaction coupling, on gauge

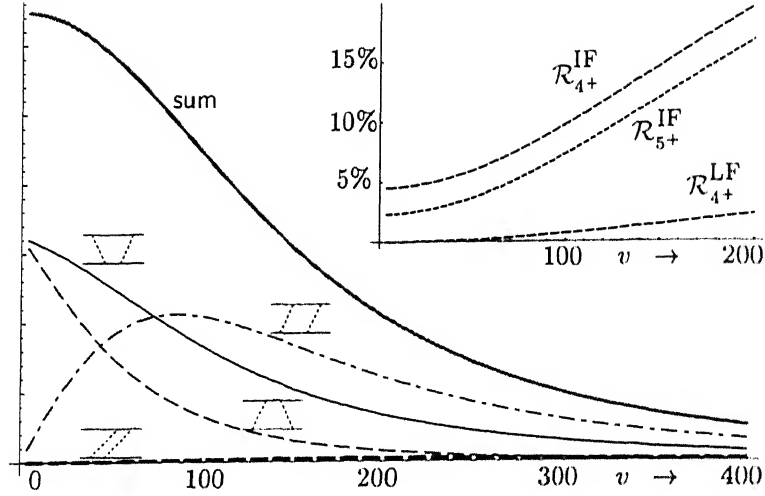


Figure 3:

coupling and on Yukawa coupling constants. So we have to solve the whole system of coupled differential equations. This can be done numerically with the help of computer. The result of calculation for  $\Lambda = 10^{15}$  GeV is presented in the Fig.3.

This is so to say the phase diagram in the plane  $m_t$  and  $m_H$ . Allowed region is located between two curves, the lower region corresponds to unstable vacuum and for the parameters in the upper region Landau pole appears at the scale lower than  $\Lambda = 10^{15}$  GeV. For experimental value of  $m_t \simeq 175$  GeV the allowed region for  $m_H$  is very strong

$$170 \text{ GeV} < m_H < 190 \text{ GeV} \quad (\Lambda = 10^{15} \text{ GeV}).$$

For  $\Lambda \simeq 10^5$  GeV the upper bound is much weaker.

## 5 Conclusions.

The Quantum Field Theory has been developed as the Fundamental Theory of Nature. Later on it was realized that the renormalizable QFT describe only the low-energy fluctuations below cut-off and do not pretend to describe physics beyond this cut-off. As we have learn from the recent development of theory it is not excluded that at short distances the fundamental theory is the string theory. Thus it may happen that QFT is not the final theory of nature. Nevertheless QFT as an effective theory that governs low-energy phenomena will remain with us forever.

### Acknowledgements

I am grateful to my numerous friends and colleagues from whom I have got the main lessons in Quantum Field Theory.

## REFERENCES

- Steven Weinberg, "The Quantum Theory of Fields:Foundations", Vol.1, Cambridge Univ.Pr.,1995  
and "Quantum Theory of Fields:Modern Applications",Vol.2, Cambridge Univ.Pr.,1996 .
- Michael E.Peskin, Daniel V.Schroeder, "An Introduction to Quantum Field Theory", Addison-  
Wesley Pub.Co.,1995
- Martinus Veltman, "Diagrammatica:The Path to Feynman Rules", Cambridge Univ.Pr.,1994
- Lev B.Okun, "Leptons and Quarks" ,Amsterdam,North-Holland,1982

# 4. Broken Reflection Symmetries

P.K. Kabir \*

Beams Physics Laboratory, University of Virginia, Charlottesville, VA USA

## Abstract

Weak interactions do not possess the symmetries with respect to discrete transformations of space-inversion  $P$ , particle-antiparticle exchange  $C$ , and motion-reversal  $T$ , which characterize the strong interactions. Nonetheless, all interactions appear to be invariant with respect to the overall symmetry corresponding to the simultaneous action  $TCP$  of all three inversions. The discovery of the broken symmetries - which are yet to be understood - is reviewed, together with prospects for their future study.

## 1 Introduction

More than forty years have elapsed since the startling discovery[1,2], in experiments suggested by Lee and Yang[3], that the phenomena of nuclear beta-decay, and the closely related processes of pi- and mu-meson decay, clearly distinguish between left and right. It was found that the mirror- image of the distribution of decay electrons from a sample of polarized Cobalt-60 nuclei, or of polarized mu-mesons, is distinctly different from the observed distribution. Previously, the assumption - considered by many to be self-evident - that the laws of physics should not distinguish between left and right in any way, had been shown by Wigner[4] to lead to the conservation of a quantum-mechanical entity called parity,- which is the eigenvalue of the space-inversion operator  $P$ ,- which explained previously unexplained regularities of atomic, and later also nuclear, spectroscopy. The unexpected discovery, that the invariance of physical laws under space-inversion is not an universal property of all interactions, was followed by further surprises. On the basis of theoretical arguments advanced by Lueders and Pauli[5,6], Lee, Oehme and Yang[7] could conclude that the large space-asymmetries found in beta-decay and in pi-mu decays would require a simultaneous breakdown of symmetry with respect to particle-antiparticle exchange  $C$ , which had been an article of faith since the development of relativistic quantum theory by Dirac. This permits the aesthetically appealing hypothesis[8] of  $CP$ -invariance, expressed elegantly by Landau through the statement that physics seen through the mirror shows the physics of anti- matter. The failure of  $C$ -symmetry in pi-mu decays was subsequently confirmed directly, when it was shown that pions of opposite charges yield oppositely polarized muons. Lee, Oehme and Yang had already noted that equality of masses and lifetimes for particles and antiparticles could not be invoked as evidence for  $C$ -invariance because those are assured[7] by the more general requirement of  $TCP$ -invariance alone. That analysis proved to be highly prescient because it was discovered that the attractive hypothesis of  $CP$ -invariance, which accommodates  $P$ - and  $C$ -noninvariance while preserving complete symmetry between matter and antimatter, also could not be exact. The  $TCP$  theorem, discussed by Lueders and Pauli, then requires that  $T$ -invariance must also fail, a prediction which appears to have been confirmed recently. The hypothesis of  $TCP$ -invariance should, of course, also be subjected to experimental test. Further arguments to do so are based on recent speculations that the correct description of fundamental particle interactions require additional space-time dimensions beyond the usual four of Minkowski space-time. If these additional dimensions are "rolled up" within a small enough region at each point of the usual 4-dimensional space-time, their existence would not be detected until one probed the corresponding scales, just as the finite size and internal structure

---

\*Email : pkk@virginia.edu

of atomic nuclei does not play any significant role in atomic and molecular phenomena. But these additional dimensions could introduce apparent non-locality in the usual 4-dimensional description, and thereby invalidate the premises of the Lueders-Pauli theorem. Refined tests of TCP-invariance are possible by studying decays of neutral K-mesons, both singly and in correlated pairs as from  $\phi$ -decay. Thus far, there has been no indication of any significant deviation from TCP-symmetry.

## 2 Space-Inversion

The deviation from space-reflection symmetry in nuclear beta-decays was first demonstrated by Wu, Ambler, Hayward, Hoppes and Hudson [1] who found that the beta-particles ( negative electrons ) emitted by a polarized source of Cobalt-60 nuclei are distributed anisotropically with respect to the nuclear polarization, with a strong preference to be emitted in a direction opposite to the nuclear spins. A closely related effect, first predicted by Landau[8], which was also later confirmed, is that these electrons should have their spins preferentially oriented against their direction of motion. If beta-decay interactions were reflection-invariant, oppositely-spinning electrons should appear as frequently, and there should not be any net polarization of the emitted electrons. Therefore, the detection of longitudinal polarization of beta-decay electrons is a clear demonstration of the breakdown of reflection-symmetry in nuclear beta-decay.

Even before experiments revealed that beta-decays show large departures from mirror-symmetry, seen also in the pi-mu decay sequence, it had been suggested[9,8,10] that these asymmetries might be linked to special properties of the neutrino, which is emitted in each of these processes. The electrically uncharged neutrino is similar to the electron in all other respects except that its mass is much smaller- so small in fact that the measured value is not yet distinguishable from zero. Its existence was conjectured by Pauli in 1930, to account for some puzzling features of beta-radioactivity, and incorporated by Fermi in his very successful phenomenological theory[11] of beta-decay. It is a property of massless spinning particles, and of such particles alone, that they can occur in states with a single spin orientation only: either directly along, or opposite to their direction of motion. By postulating that an emitted neutrino in beta-decay is always in one of these states[12]- which clearly has a handedness, depending on its spin-orientation - one would have a natural explanation of why beta-decay prefers one handedness over another. The observed departure from reflection-symmetry appeared to conform exactly to the hypothesis that the (anti)neutrinos emitted in beta-decay are purely right-handed. The corresponding neutrinos emitted in K-capture should then be purely left-handed. Although the neutrino itself had been directly detected only two years earlier, this prediction was strikingly confirmed in an experimental *tour de force*[13] within a year. Furthermore, the requirement that neutrinos are emitted only in such "chiral" states would assure that neutrinos remained strictly massless, even allowing for virtual higher-order processes which, in Fermi's theory, led formally to infinitely divergent mass-shifts. Application of the condition of masslessness to the process of muon decay, which involves the emission of a pair of neutrinos, also led to the correct prediction of the energy-distribution of the decay electrons.

Despite these successes, attribution of reflection-noninvariance to the special properties of the neutrino was unlikely to provide a complete explanation. One would still have to explain the failure of P-conservation in processes without neutrinos, in particular the coexisting 2-pion and 3-pion decays of K-mesons, the original problem which led to the proposal of P-nonconservation. Feinberg[14] had already noted that the successful predictions for muon decay were in fact "accidental" because even the expectation of parity nonconservation in that process was not required unless one recognized the existence of two distinct neutrinos, a fact which was not invoked by any of the authors of Refs.[9,8,10]. We now know that the correct form of the Fermi Interaction is obtained by requiring each of the participating fermions to interact through the same chiral projection[15,16], and the neutrino does not have a privileged role. Indeed the current view is that neutrinos, which are believed to occur in at least three varieties, probably have small masses, which can be invoked to explain various outstanding problems[17].

The Fermi theory is now known to be the low-energy limit of a theory in which charged vector mesons, with masses almost a hundred times greater than the proton mass, are exchanged

between charge-changing currents formed from chiral fermions. Why a particular chirality is chosen, and why the participating fermion fields are certain specific superpositions of the observed fermions ( with well-defined masses ) are important questions for which no satisfactory explanation is available at present. The currently favoured description, that these are manifestations of spontaneous symmetry-breaking, accomodates the observed parameters while renouncing any attempt to explain them.

### 3 The Failed Hypothesis of CP Symmetry

From the viewpoint of symmetry, the hypothesis of CP-invariance, advocated most forcefully by Landau[8], offered an extremely attractive solution to the problem posed by the discovery that beta-decay and other weak interactions distinguish between left and right. The difference between the beta-decay distribution from a polarized Cobalt-60 source, and its mirror-image, makes no distinction between left and right if there is exact CP-symmetry. Although beta-decay may not appear to be mirror-symmetric, CP-invariance would restore mirror-symmetry in a larger sense. By requiring that the mirror-image of any physical process should represent the course of the related phenomenon where each particle is replaced by its corresponding anti-particle, the symmetry of space would be completely restored and any observed difference would reflect only the reciprocal difference in behaviour between particles and their corresponding antiparticles. It would *not* be possible to give an absolute definition of left or right if one did not know whether one was dealing with particles or antiparticles. Conversely, one would not be able to tell whether one was observing a group of particles or their corresponding antiparticles, solely by observing their interactions. From 1957 to 1964, all deviations from reflection-symmetry observed in weak decay processes appeared to conform to the broader requirement of CP-invariance. There was an additional theoretical argument in favour of CP-symmetry. According to the theoretical expectation of TCP-invariance [ see next Section], CP-invariance would necessarily require T-invariance. Thus CP- symmetry would not only guarantee the symmetry of space and the absolute equality of left and right, and of matter and antimatter, it would also require the validity of exact T-invariance. Wigner had shown[4] that many regularities,- notably the degeneracy discovered by Kramers for energy eigenstates of electrons in potential fields of arbitrary complexity - could be understood most simply under this hypothesis. Furthermore, the discovery of parity non-conservation invalidated the usual argument for the non-occurrence of static electric dipole moments for atoms and molecules, but Landau showed[8] that T-invariance was sufficient to forbid their occurrence, whether parity was conserved or not. Consequently, the failure thus far to detect electric dipole moments - in the case of the neutron at a scale  $10^{-13}$  smaller than the known extension of its charge and magnetization - could be considered as support for this hypothesis.

Therefore, it came as another great surprise when, in 1964, experiments revealed that CP could not be sustained as an exact symmetry of Nature. To explain the discovery, we must first make a brief digression about neutral K-mesons. Among the first "strange" particles to be discovered was a neutral particle with a mass of about  $500\text{MeV}/c^2$  and a lifetime of about  $10^{-10}\text{sec}$ , which decayed into a pair of pi- mesons. To account for the slowness of its decay, Gell-Mann and Nishijima independently proposed the notion of "strangeness", an additive property similar to electric charge, appropriately assigned to the  $K^0$  and other "strange" particles, which is conserved in the strong interactions which produce K-mesons, but not in the weak interactions which lead to their decay. In the Gell-Mann-Nishijima scheme, the antiparticle  $\bar{K}^0$  must have strangeness opposite to that of the  $K^0$  and therefore be a physically distinct state; TCP-invariance would require that it be degenerate with  $K^0$ . Since, by hypothesis, strangeness is not conserved in weak decay processes, the decay of a  $K^0$  would be accompanied by mixing with the  $\bar{K}^0$  state:  $K^0$  decay cannot be correctly described without taking account of the degenerate  $\bar{K}^0$  state with which it mixes.

Gell-Mann and Pais[18] showed that decays of neutral K-mesons are more conveniently described in terms of the coherent superpositions ( particle symbols represent the corresponding quantum states):

$$K_{\pm} = (K^0 \pm \bar{K}^0)/\sqrt{2} \quad (1)$$

which are eigenstates of CP, with eigenvalues  $\pm 1$ , respectively, if CP is an exact symmetry. Invariance of all interactions under CP-transformation would require CP to be conserved. Then the CP-even eigenstate  $K_+$  could decay only to CP-even eigenstates while  $K_-$  would correspondingly decay only to CP-odd eigenstates. Production of a  $K^0$  in a nuclear collision should be viewed as creation of the coherent superposition:

$$K^0 = (K_+ + K_-)/\sqrt{2}; \quad (2)$$

$\bar{K}^0$  production would likewise correspond to creation of a similar superposition, in which the phase of  $K_-$  is reversed. Since a  $2\pi$  s-state is necessarily a CP-even state, the  $\theta^0 \rightarrow \pi^+\pi^-$  mode of  $K^0$ 's originally observed must represent decays of the  $K_+$  component; the  $K_-$  component which, according to Eq.(2), is produced with equal probability as  $K_+$ , can *not* decay via  $2\pi$  modes and would therefore presumably decay more slowly via 3-body modes. Confirmation of the occurrence of such longer-lived neutral K-mesons, a spectacular prediction of quantum mechanics, with a mean life of about  $6.10^{-8}\text{sec}$  decaying via 3-body decay channels, was one of the first triumphs of the particle-mixture theory.

To the surprise of almost everyone, Christenson, Cronin, Fitch and Turlay discovered[19] that about one in every 500 long-lived K-mesons also decays into a pair of  $\pi$ -mesons, contradicting the theoretical expectation from the hypothesis of CP-invariance. Many interpretations, which sought to explain this observation without giving up CP-symmetry, had to be abandoned after it was found[20] that the  $2\pi$  state arising from decay of the long-lived state  $K_L$  is quantum-mechanically identical with the one resulting from decay of the short-lived state  $K_S$ , because it interferes constructively with it. This shows that the long-lived state  $K_L$  contains a CP-even admixture, as conjectured by the original authors, or that the decay permits a CP-odd initial state to make a transition to a CP-even final state. In either case, CP-conservation must be abandoned. The most direct and striking demonstration of the breakdown of CP-symmetry comes from a comparison of the time-distribution of  $2\pi$  decays from a sample of initial  $K^0$ 's with that from a sample of initial  $\bar{K}^0$ 's. Eq.(2) shows that an initial  $K^0$  state is an equal superposition of  $K_+$  and  $K_-$

(which, in lowest approximation, may be identified with the short- and long-lived kaons states  $K_S$  and  $K_L$ , respectively). Since the  $K_S$  decays with a lifetime  $\tau_S$  much shorter than that of  $K_L$ , the amplitude of  $K_+$  relative to  $K_-$ , will rapidly reduce to a level such that the relatively feeble ( $\eta$  times smaller)  $K_L \rightarrow 2\pi$  decay amplitude becomes comparable to the  $K_S$  contribution. This should happen at  $\tau \approx 2\tau_S |\log_e \eta| \approx 12\tau_S$  for  $|\eta| = 2.10^{-3}$ . For such times, the replacement of  $K^0$  by  $\bar{K}^0$ , which reverses the relative sign between the two terms in Eq.(2), should result in a dramatic change of the decay rate as constructive interference replaces destructive interference and *viceversa*. Such a clearcut difference between the time-distributions of  $K^0 \rightarrow \pi^+\pi^-$  and  $\bar{K}^0 \rightarrow \pi^+\pi^-$  decays is indeed observed.

Because of the degeneracy between  $K^0$  and  $\bar{K}^0$  states, an extremely weak CP-noninvariant interaction suffices to provide the small CP-even admixture  $\epsilon$  in the long-lived kaon state:

$$K_L \propto K_- + \epsilon K_+ \quad (3)$$

which is required to account for the observed departures from CP-symmetry. Until very recently, it appeared that the parameter  $\epsilon$  (whose phase, under reasonable assumptions, is fixed by other measured parameters) was the only known measure of CP-noninvariance, which had not been seen anywhere outside the neutral K-meson system. The KTeV group at Fermilab has now confirmed an earlier report[21] from CERN that the  $\pi^+\pi^-/\pi^0\pi^0$  ratio in  $K_L$  decays is 1.7% higher than the corresponding ratio for  $K_S$  decays. If the admixture  $\epsilon$  of the CP-even component  $K_+$  in the long-lived neutral kaon state  $K_L$  were the only cause of the observed CP-noninvariance, the  $\pi^+\pi^-/\pi^0\pi^0$  ratio in  $K_L$  decays should be the same as that for  $K_S$ . Thus, there must be a "direct" contribution to  $2\pi$  decays of neutral K-mesons, which also rules out explanations which attribute the observed CP-asymmetry to external effects, such as the influence of hypothetical long-range fields which affect  $K^0$  and  $\bar{K}^0$  differently, or interactions with ambient CP-asymmetric media, such as a neutrino "sea". The detection of "direct" CP-noninvariant interactions greatly



encourages the search for CP-asymmetric effects outside the neutral kaon system. Such searches have been conducted extensively since the original discovery but, in the absence of theoretical guidance, these could not be more than gropings in the dark, and did not lead to any further evidence for CP-noninvariance. As long as  $\epsilon$  sufficed to describe the observed CP-noninvariance in neutral kaon decays, no further CP-noninvariant effects could be predicted with any assurance. Nonetheless, the discovery of new kinds of strangeness, now called "charm" and "beauty", led to the prediction and discovery of mesons bearing these attributes. Among these were more massive homologues of the  $K^0$ , called  $D^0$  and  $B^0$ , respectively, whose transformations and decays should be similar to those of neutral kaons, and be described by a similar formalism. In particular, mixing of  $B^0$  with  $\bar{B}^0$  appears to be well-established[22] and the currently-favoured parametrization of weak-interaction currents suggests that relatively large CP-asymmetries could occur in B-decays. In the Standard Model, which satisfactorily describes most high-energy phenomena, CP-noninvariance is accommodated through mixing of the charged chiral currents which generate weak interactions. For currents formed from three pairs of quark fields, Kobayashi and Maskawa found[23] that the most general unitary mixing involves a complex phase angle, whose presence leads to CP-noninvariant effects. Jarlskog[24] and others showed that, if this is the only source of CP-noninvariance, all such observables are proportional to a single phase-invariant parameter  $J$  characteristic of the mixing-matrix which describes the weak currents. In principle, measurement of the CP-violating parameter  $\epsilon$  in neutral K-meson decay should then fix  $J$ ; all other CP-asymmetric effects in  $B^0$  and other decays would then be predicted. Unfortunately, present inability to reliably calculate effects of strong interactions, - arising from Quantum Chromodynamics in the Standard Model - prevents the full implementation of this programme. Within the limits of such uncertainties, there are suggestions[22] of CP-nonconserving effects significantly larger than in  $K^0$ -decays. Dedicated B-factories - which are  $e^+e^-$  colliders designed to produce copious  $B - \bar{B}$  pairs - have been constructed to search for these processes, and results may be expected soon. A  $\phi$ -factory called  $D\Lambda\Phi NF$  is already operating at Frascati;  $\phi \rightarrow K_L K_S$  decays provide a unique means to study pure  $K_S$  samples, whose CP-violating decays have not yet been measured.

## 4 TCP-Invariance

Even before the discovery that weak interactions do not respect the reflection-symmetries which characterize strong and electromagnetic interactions, several authors[25,5,6] had noted the existence of a more general kind of reflection-symmetry which appears to have a stronger theoretical foundation. Named "strong reflection" by Schwinger, it corresponds to simultaneous inversion of space- and time-coordinates. But for the fact that the time-coordinate is singled out by its opposite sign in the relativistic metric, such a transformation could be accomplished by a pure rotation in the even-dimensional Minkowski space, and thus be an element of the class of *proper* Lorentz transformations. By extending the concept of Lorentz-invariance to include invariance with respect to Lorentz transformations for complex values of the relative velocity ( which determines the rotation angle ), the requirement of Lorentz-invariance would include invariance under strong reflection. Under strong reflection, energy (which transforms like the time-coordinate) must change sign but, in relativity theory, negative energies correspond to negative inertia and therefore make no sense. Following Dirac and Feynman, the problem is avoided by reinterpreting these states of negative energy, found by applying strong reflection to particle-states of positive energy, as corresponding positive-energy states of *anti*-particles. Thus, strong reflection can be given an unambiguous meaning as the operation of CPT or TCP ( the order of the factors is irrelevant since all the inversions are taken to act together ), viz. inversion of space- and time-coordinates *simultaneously* with particle-antiparticle conjugation. Local Lorentz-invariant quantum field theories can readily be constructed which do *not* have symmetry with respect to one or more of the discrete transformations P, C, or T but all of them would necessarily be invariant with respect to the combined operation of strong reflection  $\Theta \equiv \text{TCP}$ . The existence of anti-particles - predicted by Dirac - is a consequence not of C-invariance ( which may or may not hold in particular cases ) of relativistic quantum field theory, as originally supposed, but of the more general requirement of

TCP-invariance.

In addition to the requirement that particles and antiparticles have equal masses and lifetimes, TCP-invariance requires them to have identical electromagnetic properties, apart from obvious changes of sign. Thus far, all measurements are consistent with these conditions. By far the most stringent test comes from a comparison of  $K^0$  and  $\bar{K}^0$  masses, by applying a relation given by Bell[26]. Any difference between the masses or decay-widths of  $K^0$  and  $\bar{K}^0$  would be reflected in the complex mass-difference between  $K_L$  and  $K_S$  and the composition of the  $K_L$  and  $K_S$  states through:

$$\Lambda - \bar{\Lambda} = (\lambda_L - \lambda_S)(\epsilon_S - \epsilon_L), \quad (4)$$

if one neglects terms higher than quadratic in the  $\epsilon$ 's. If one does not assume TCP-invariance,  $\epsilon_L$  replaces the  $\epsilon$  in Eq.(3), while  $\epsilon_S$  is defined by a corresponding equation for  $K_S$ . Using the best available data, the Particle Data Group reports[27] a limit for the LHS which cannot exceed  $10^{-18}$  times the kaon mass! This is by far the best test of TCP-invariance available at present.

## 5 Question of T-Invariance

According to the hypothesis of TCP-invariance, any deviation from CP-symmetry must be accompanied by a compensating departure from T-symmetry. Therefore, it is important both as a check of the TCP-theorem and on general principles, to search for departures from T-invariance which, in some way or other, must correspond to deviations from reciprocity.

As long as all known departures from CP-symmetry are restricted to the neutral K-meson system, where they can be described[28] by the parameter  $\epsilon$ , it follows that the only T-noninvariant effects which can be predicted with assurance are those dependent on the same parameter. In the TCP-invariant description,  $K^0 \rightleftharpoons \bar{K}^0$  transitions do not satisfy reciprocity[29], and a corresponding asymmetry has been found[30] between leptonic decays of  $K^0$  and  $\bar{K}^0$  with the expected sign and magnitude. The observed effect does *not* constitute a direct test of reciprocity, thus one cannot assert unambiguously[31] that T-noninvariance has been directly demonstrated, but the cost of defending T-invariance would necessarily demand the sacrifice of TCP-symmetry. Fortunately, this question can be settled by direct experimental test, and an answer should be available soon.

## 6 Conclusions

Departures from symmetry under space-reflection P, and under particle-antiparticle exchange C, characterize the weak interactions. The hypothesis of CP-invariance, or symmetry under combined inversion, would restore the symmetry of space, and the equivalence of particles and antiparticles, just as Pasteur's discovery of stereoisomerism restored the symmetry between right and left in an earlier era. The conclusive discovery of CP-noninvariance in neutral K-meson decays demolishes that possibility. The fact that the deviation from CP-symmetry is small, and observed thus far only in neutral kaon decays, may be related to the possibility that CP-noninvariant interactions require the participation of much more massive particles. Such particles could contribute only virtually, and relatively weakly by the rules of quantum theory, in the decay of lighter particles. Their effects are detectable in decays of neutral kaons because of the special circumstance of  $K^0 - \bar{K}^0$  degeneracy, required by TCP-invariance. The corresponding deviation from T-invariance, demanded by CP-noninvariance, appears to be confirmed. Larger CP-noninvariant effects may occur in decays of the more massive B mesons. Detailed studies of neutral kaon mixing, both for individual particles and for kaon pairs, could provide improved tests of TCP.

From a fundamental point of view, the broken mirror-symmetries, with respect to P, C and T, remain completely mysterious. Thus far, there is no evidence of any deviation from symmetry under the combined transformation TCP of "strong reflection", whose basis in relativistic quantum field theory furnishes the only available explanation for the occurrence of antiparticles.

## References

- [1] Wu, Ambler, Hayward, Hoppes and Hudson, Phys. Rev. 105,1413 (1957).
- [2] Garwin, Lederman and Weinreich, Phys. Rev. 105,1415 (1957).
- [3] T.D. Lee and C.N. Yang, Phys. Rev. 104,256 (1956).
- [4] E.P. Wigner, Group Theory and Applications to Quantum Mechanics, Academic Press, New York, 1959.
- [5] G.Lueders, Kgl. Dan. Vid. Selskab,Mat-Fys. Med. 28,No. 5 (1954).
- [6] W. Pauli,in "Niels Bohr and the Development of Physics", Pergamon, London, 1955.
- [7] Lee, Oehme and Yang, Phys. Rev. 106,340 (1957).
- [8] L.D. Landau, Nucl. Phys. 3,127 (1957).
- [9] A. Salam, Nuovo Cimento 5, 299 (1957).
- [10] T.D. Lee and C.N. Yang, Phys. Rev. 105, 1671 (1957).
- [11] E. Fermi, Zeits. f. Phys. 88,161 (1934).
- [12] Such a theory was proposed by Weyl in the early days of quantum theory, but was rejected by Pauli precisely because it was not consistent with the requirement of reflection-invariance.
- [13] Goldhaber, Grodzins and Sunyar, Phys. Rev. 109,1015(1958).
- [14] G. Feinberg, private communication, 1957.
- [15] E.C.G. Sudarshan and R.E. Marshak, Proc. Padua-Venice Conf. 1957.
- [16] R.P. Feynman and M. Gell-Mann, Phys. Rev. 109,193 (1958).
- [17] Among these are the solar neutrino deficit, viz. the fact that the rate of neutrino reactions observed on Earth, induced by neutrinos produced by nuclear reactions in the solar interior, is about one-third of the expected rate, and the anomalous ratio of high-energy muons to electrons produced deep underground, presumably by neutrinos from decay of mesons created by cosmic-rays collisions with the atmosphere.
- [18] M. Gell-Mann and A. Pais, Phys. Rev. 97,1387 (1955).
- [19] Christenson, Cronin, Fitch and Turlay, Phys. Rev. Lett.13,138 (1964).
- [20] Fitch, Ross, Russ, and Vernon, Phys. Rev. Lett. 15,73 (1965).
- [21] NA31 Group. G.D.Barr et al. Phys. Lett.B317, 233 (1993).
- [22] See Y. Nir and H. Quinn, Ann. Rev. Nucl.and Part.Sc 42,211 (1992).
- [23] M. Kobayashi and T. Maskawa, Prog. Theor. Phys. 49,652 (1973).
- [24] C.Jarlskog, Phys. Rev. Lett.55,1039 (1985).
- [25] J. Schwinger, Phys. Rev. 82,914 (1951).
- [26] J.S. Bell, Proc. Oxford Intl. Conf. on El.Particles 1965, R.G.Moorhouse et al. eds. Rutherford High Energy Laboratory, 1966.
- [27] Particle Data Group. Eur.Phys.J. C3, 1 (1998).

- [28] The additional parameter required to describe the effect reported in Ref.21 is too small to substantially affect the following conclusion.
- [29] P.K. Kabir, Phys. Rev. D2,540 (1970).
- [30] CPLEAR Group. A. Angelopoulos et al. Phys.Lett.B444,43 (1998).
- [31] P.K. Kabir, Phys. Lett. B459,335 (1999).

# 5. Dynamics Of Symmetry Breaking Out Of Equilibrium: From Condensed Matter To QCD And The Early Universe\*

D. Boyanovsky<sup>(a,b)</sup> and H.J. de Vega<sup>(b,a)</sup>

Department of Physics and Astronomy,  
University of Pittsburgh, Pittsburgh, PA 15260 USA

(b) LPTHEL<sup>†</sup>

Université Pierre et Marie Curie (Paris VI) et Denis Diderot (Paris VII),  
Tour 16, 1er. étage, 4, Place Jussieu 75252 Paris, Cedex 05, France

## Abstract

The dynamics of symmetry breaking during out of equilibrium phase transitions is a topic of great importance in many disciplines, from condensed matter to particle physics and early Universe cosmology with definite experimental and observational impact. In these notes we provide a summary of the relevant aspects of the dynamics of symmetry breaking in many different fields with emphasis on the experimental realizations. In condensed matter we address the dynamics of phase ordering, the emergence of condensates, coarsening and dynamical scaling. In QCD the possibility of disoriented chiral condensates of pions emerging during a strongly out of equilibrium phase transition is discussed. Finally we elaborate on the dynamics of phase ordering in phase transitions in the Early Universe, in particular the emergence of condensates and scaling in FRW cosmologies. We mention some experimental efforts in different fields that study this wide ranging phenomena and offer a quantitative theoretical description both at the phenomenological level in condensed matter introducing the scaling hypothesis as well as at a microscopic level in quantum field theories. The emergence of semiclassical condensates and a dynamical length scale is shown in detail, in quantum field theory this length scale is constrained by causality.

The large  $N$  limit provides a natural bridge to compare the solutions in the different settings and to establish similarities and differences. 11.10.-z;11.15.Pg;11.30.Qc

## 1 Phase Ordering Dynamics: an interdisciplinary fascinating problem

The dynamics of non-equilibrium phase transitions and the ordering process that occurs until the system reaches a broken symmetry equilibrium state play an important role in many different areas. In condensed matter physics binary fluids, ferromagnets, superfluids, and liquid crystals, to name a few, are examples of systems in which the dynamics of phase transitions out of equilibrium are studied experimentally.

Experiments in these systems have provided a solid basis for the description of the dynamics of phase ordering: in binary fluids or alloys upon a sudden temperature drop below the critical temperature, the two fluids begin to separate, regions of different fluid concentrations are separated by *domain walls*. In superfluids, rapid cooling leads to a network of vortices and in liquid crystals to many different topological defects.

---

\*Email: BOYAN@vsml.cis.pitt.edu

<sup>†</sup>laboratoire Associé au CNRS UMR 7589.

Current and future measurements of Cosmic Microwave Background anisotropies as well as the formation of large scale structures in the universe provide distinct evidence for phase transitions during inflation and after [1]. At even lower energies, available with current and forthcoming accelerators, the Relativistic Heavy Ion Collider (RHIC) at Brookhaven and the Large Hadron Collider (LHC) at Cern the phase transitions predicted by the theory of strong interactions, Quantum Chromodynamics (QCD) could occur out of equilibrium via the formation of coherent condensates of low energy Pions.

These phase transitions and the associated processes which often take place out of equilibrium will be an experimental telltale of the chiral phase transition of QCD[5].

Whereas the GUT phase transition took place when the Universe was about  $10^{-35}$  seconds old and the temperature about  $10^{23}K$ , and the EW phase transition occurred when the Universe was  $10^{-12}$  seconds old and with a temperature  $10^{15}K$ , the QCD phase transition took place at about  $10^{-5}$  seconds after the Big Bang, when the temperature was a mere  $10^{12}K$ . This temperature range will be probed at RHIC and LHC within the next very few years. The basic problem of describing the process of phase ordering, the competition between different broken symmetry states and the formation and evolution of condensates on the way towards reaching equilibrium is common to all of these situations and fields. The tools, however, are necessarily very different: whereas ferromagnets, binary fluids or alloys etc, can be described via a phenomenological (stochastic) classical description, certainly in quantum field theory a microscopic formulation must be provided. In these lectures we describe a program to include ideas from condensed matter to the realm of quantum field theory, to describe the non-equilibrium dynamics of symmetry breaking and the process of phase separation and phase ordering on a range of time and spatial scales of unprecedented resolution (for example in QCD the time scales  $\leq 10^{-23}$  seconds, spatial scales  $\leq 10^{-15}$  meters, in cosmology the time scales are of order  $10^{-32}$  seconds and spatial scales smaller than  $10^{-44}$  meters) that require a full quantum field theoretical description.

We begin the excursion into these timely fields by first providing a brief quantitative description of the relevant setting and whenever possible the experimental situation associated with them in three main areas: Condensed Matter, Ultrarelativistic Heavy Ion Collisions and Early Universe Cosmology. This qualitative discussion will be followed by a more quantitative description of some of the main theoretical ideas, techniques and tools.

## 1.1 Condensed Matter:

A description of phase transitions and critical phenomena in equilibrium begins by recognizing an *order parameter* which is a thermodynamic ensemble average of a macroscopic variable that determines the different macroscopic states of the system. For example in ferromagnets the order parameter is the average magnetization, above a critical temperature it vanishes and it is non-zero below the critical temperature, in superfluids is the condensate density, in superconductors the density of Cooper pairs, etc.[6]. Phase transitions *in equilibrium* are fairly well understood and described by the theory of critical phenomena[6] which combined with the renormalization group provides a very successful description of phase transitions. The theory of critical phenomena and the renormalization group provide a very robust description of *universality classes*: many systems that are very different behave similarly near critical points, these universality classes are divided by for example the dimensionality of the order parameter, the dimensionality of space, and the symmetries of the underlying microscopic Hamiltonian. An important concept in critical phenomena is the correlation length, take for example a spin one-half ferromagnet, the microscopic Hamiltonian has an up-down symmetry, the energy remains the same if all spins are flipped. Focus at a particular point of the sample where the spin is up. The correlation length is the distance over which the spins are correlated, i.e. the distance from this up-spin over which the neighboring spins are also up. As the critical temperature is reached from above this correlation length grows reaching a macroscopic size (diverging) at the critical temperature. As the system is cooled below the critical temperature a phase transition occurs: there appears a net overall magnetization and at low temperatures all spins are either up or down, the up-down symmetry is spontaneously broken[6]. This phase transition occurs in equilibrium when the microscopic relaxation time scales

are shorter than the time scale of cooling the system, thus at all times the system is in *local thermodynamic equilibrium*. At very high temperatures typically the disordered phase prevails, all spins are oriented at random and the average magnetization vanishes. As the critical temperature is reached regions of correlated spins appear and become of macroscopic size as the correlation length diverges and the spin system begins to order. In this region the thermodynamic quantities become insensitive to the short distance details such as crystalline lattices, lattice spacing and the nature and strength of the interaction between spins as the physics is determined by the correlation of spins over large distances. Near the critical point, the short distance length scales are irrelevant for macroscopic phenomena and long-wavelength physics is completely determined by the correlation length  $\xi(T)$ . Macroscopic thermodynamic quantities and susceptibilities near the critical temperature *only* depend on the length scale  $\xi(T)$ .

This is the basis of the *static* scaling hypothesis which is confirmed experimentally in a wide variety of systems and is theoretically supported by the renormalization group approach to critical phenomena[6]. The *static* critical phenomena associated with second order phase transitions that occur in local thermodynamic equilibrium is fairly well understood via the renormalization group (and other alternative approaches)[6].

Consider the alternative scenario in which a ferromagnet is held at very high temperature in the disordered phase and suddenly it is cooled below the critical temperature on time scales shorter than those associated with relaxational phenomena. Now the spin system must evolve towards the ordered phase far away from equilibrium. Unlike the case of static (local thermodynamic equilibrium) critical phenomena, the case of out of equilibrium phase transitions require a novel set of ideas and tools to describe the *dynamics* of the process of phase ordering.

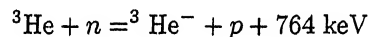
There is now a large body of theoretical and experimental work in phase ordering dynamics in condensed matter systems[7]-[10]. Although ultimately the tools to study similar questions in quantum mechanical many body systems will be different, the main physical features to describe are basically the same: as the system cools down suddenly below the critical temperature correlated regions (of spins in a ferromagnet or of condensate in a Bose superfluid) begin to form. These correlated regions are separated by 'walls' or other structures. Inside these regions an ordered phase exists which eventually grows in time to become macroscopic in size. Before attempting to describe the manner in which a given system orders after being cooled through a phase transition an understanding of the relevant time scales is required. Two important time scales determine if the transition occurs in or out of equilibrium: the relaxation time of long wavelength fluctuations (since these are the ones that order)  $\tau_{rel}(k)$  and the inverse of the cooling rate  $t_{cool} = T(t)/\dot{T}(t)$ . If  $\tau_{rel}(k) \ll t_{cool}$  then these wavelengths are in local thermodynamical equilibrium (LTE), but if  $\tau_{rel}(k) \gg t_{cool}$  these wavelengths fall out of LTE and freeze out, for these the phase transition occurs in a quenched manner. These modes do not have time to adjust locally to the temperature change and for them the transition from a high temperature phase to a low temperature one occur instantaneously. This description was presented by Zurek[11] analysing the emergence of defect networks after a quenched phase transition. Whereas the short wavelength modes are rapidly thermalized (typically by collisions) the long-wavelength modes with  $k \ll 1/\xi(T)$  with  $\xi(T)$  the correlation length (in the disordered phase) become *critically slowed down* i.e. their relaxation time becomes extremely long near the critical point. As  $T \rightarrow T_c^+$  the long wavelength modes relax very slowly, they fall out of LTE and any finite cooling rate causes them to undergo a 'quenched' non-equilibrium phase transition. As the system is quenched from  $T > T_c$  (disordered phase) to  $T < T_c$  (ordered phase) ordering *does not* occur instantaneously. The length scale of the ordered regions grows in time (after some initial transients) as the different broken symmetry phases compete to select the final equilibrium state. A *dynamical* length scale  $\xi(t)$  typically emerges which is interpreted as the size of the correlated regions, this dynamical correlation length grows in time to become macroscopically large[7, 8, 9, 10]. Just as in *static* critical phenomena, the emergence of this dynamical correlation length leads to the *dynamical scaling hypothesis*, that the approach to equilibrium and the kinetics of phase ordering is solely determined by this length scale.

Experiments in binary fluids for example, study the growth of these correlated regions by light scattering[12] much in the same manner as the onset of ferromagnetism is studied via neutron scattering. The growth of the domains, characterized by the dynamical length scale results in

that as a function of time the scattering of light becomes stronger for longer wavelengths i.e. smaller wave-vectors, until eventually at very long times a Bragg peak at zero momentum emerges signaling the macroscopic ordering of the system. This growth of domain structures during the dynamical process of phase ordering is referred to as ‘coarsening’[7, 8, 9, 10]. This mechanism with a clear experimental realization in condensed matter is at the heart of the Kibble-Zurek[2, 11, 13] scenario of the dynamics of symmetry breaking in cosmological phase transitions[2, 11, 13]. In this scenario a ‘network’ of defects emerges right after a phase transition that occurred strongly out of equilibrium with a density of about one defect per initial correlation length. This network eventually evolves and in computer simulations a scaling regime is observed to emerge[2].

The dynamics of phase ordering had been studied in liquid crystals whose symmetry group is rather similar to that of particle physics models. The experiments produced non-equilibrium phase transitions both by suddenly varying the pressure and the temperature (pressure and temperature quenches)[14, 15] and confirmed at least in a qualitative manner the main features described by this scenario of dynamics of symmetry breaking. More recently a new set of experiments had sought to provide a more detailed picture of the dynamics of symmetry breaking phase transitions out of equilibrium and to simulate in the laboratory what is thought to be the situation in cosmological phase transitions. Original experiments focused on studying the dynamics of phase ordering after a pressure quench in superfluid  $^4\text{He}$ [16] by measuring second sound (i.e. entropy disturbances) which only propagate in the superfluid component (the broken symmetry phase). The interpretation of results in these experiments were overshadowed by induced turbulence during the quench and spurious phase separation due to imperfections of the walls. More recently a new set of experiments were carried out that seem to lead to cleaner interpretations.

In these ingenious experiments[17] a small sample of superfluid  $^3\text{He}$ , whose order parameter has a group structure very similar to some particle physics models, was heated locally by neutron irradiation via the nuclear reaction



the energy released heats a small portion of the liquid Helium into the normal state and rapid diffusion of the quasiparticles cools this region back into the superfluid phase very rapidly, thus providing a quench from a normal (disordered) phase into the superfluid (ordered) phase. The resulting domain structure is then studied via NMR and a qualitative agreement with the picture of the symmetry breaking dynamics seem to emerge from these experiments. Thus these beautiful experiments in condensed matter provide controlled experimental framework to test the concepts associated with the dynamics of symmetry breaking.

These ideas of the emergence of correlated regions that grow in time and become macroscopic during non-equilibrium phase transitions has been recently invoked as a potential signature of the chiral phase transition in QCD, the theory of the strong interactions.

## 1.2 Chiral symmetry breaking in QCD and disoriented chiral condensates

Quantum Chromodynamics (QCD) is the theory of strong interactions, with the fundamental degrees of freedom being the quarks and gluons. Quarks, however are confined inside hadrons and do not exist as individual, isolated particles in vacuum. However there is now a wealth of theoretical evidence including very convincing lattice results that indicate that at temperatures above  $T \approx 150\text{MeV}$  quarks and gluons become free and form a quark-gluon plasma. The lattice results are supported qualitatively and quantitatively by phenomenological models[18]. In fact the evidence supports the picture of *two* phase transitions: the quark-gluon plasma or confining-deconfining phase transition in which the quarks and gluons become confined into hadrons and the chiral phase transition that leads to the low energy description in terms of pions. The low energy limit of QCD is dominated by the lightest up and down quarks  $u, d$  with masses  $m_u \approx 5\text{MeV}$ ;  $m_d \approx 7 - 10\text{MeV}$ . These mass scales are much smaller than the natural scale of QCD,  $\Lambda_{\text{QCD}} \approx 100\text{MeV}$  at which QCD becomes strongly coupled. In the limit of vanishing up and down quark masses, the



QCD Hamiltonian possesses a global *chiral symmetry* corresponding to rotating independently the right and left handed components of the spinors that describe the quark fields. This symmetry is  $SU(2)_L \otimes SU(2)_R$  and in the low energy world is spontaneously broken down to  $SU(2)_{L+R}$  with the charged and neutral pion isospin triplet being the (quasi) Goldstone bosons associated with the breakdown of this symmetry. The small mass of the pions ( $\approx 135\text{MeV}$ ), on the hadronic scale is a result of the small mass of the up and down quarks on the QCD scale, which breaks explicitly chiral symmetry. This is the chiral phase transition. The lattice results seem to indicate that the two transitions, deconfining and chiral symmetry breaking are very close in temperature and may actually happen at the same temperature[18].

Whereas the deconfining phase transition does not seem to be characterized by a natural order parameter, the chiral transition is described by the non-vanishing of the chiral condensate  $\langle \bar{q}q \rangle$  with  $\bar{q} = (\bar{u}, \bar{d})$  with  $\langle \dots \rangle$  referring to the vacuum expectation value or the thermodynamic ensemble average at finite temperature. Although this transition(s) have taken place when the Universe was at a temperature of 150MeV about  $10^{-5}$  seconds after the Big Bang, the Relativistic Heavy Ion Collider (RHIC) at Brookhaven to start operation at the end of 1999 and the Large Hadron Collider at Cern (around 2004) will probe this transitions by colliding heavy ions.

RHIC will accelerate and collide from protons up to 250 GeV and ions of up to the heaviest nuclei with collision energies of about 100 GeV per nucleon for Au nuclei. The phenomenon of nuclear transparency observed in nucleon-nucleon collisions leads to the conclusion that about half the energy of the collision is carried away by the nuclei and about half the energy is deposited in the 'central region' of the collision. Most of the baryons are carried by the receding nuclei leaving this central region almost baryon free. Estimates of the energy density in this region give  $\epsilon \approx 3 - 5\text{GeV/fm}^3$  corresponding to temperatures  $T \approx 200\text{MeV}$ . Immediately after the collision, hard scattering of quarks and gluons dominate the dynamics the gluons have mean-free paths estimated to be of order 0.5fm and the quarks 1 - 2fm (the difference is mainly due to color factors) hence after a time of the order of about 1fm/c the plasma is thermalized.

The next stage of the dynamics is described by Bjorken's hydrodynamic picture[19]. When the plasma has achieved local thermodynamic equilibrium and for wavelengths longer than the mean free paths, the plasma can be described as a strongly coupled fluid and a hydrodynamic description is suitable. The essential ingredients in a hydrodynamics description are i) the fluid is described by a *local* four velocity vector  $u^\mu = \gamma(1, \vec{v})$ ;  $u^\mu u_\mu = 1$ , the energy momentum tensor is that of an homogeneous and isotropic fluid

$$T^{\mu\nu} = (\epsilon + P)u^\mu u^\nu - Pg^{\mu\nu}$$

with  $\epsilon$ ,  $P$  the energy density and pressure respectively. The dynamics is then completely determined by conservation laws: i) baryon number, ii) energy momentum and by local thermodynamic equilibrium relations. The resulting picture of this hydrodynamic evolution is that the plasma expands and cools adiabatically and the temperature drops in time with the following law

$$T(\tau) = T_0 \left( \frac{\tau_0}{\tau} \right)^{c_s^2}$$

with  $c_s$  the adiabatic sound speed,  $T_0 \geq 200\text{MeV}$  and  $\tau_0 \approx 1\text{fm/c}$ . For a radiation dominated fluid  $c_s^2 = 1/3$ .

As the critical temperature for the chiral phase transition is reached from above the long-wavelength fluctuations of the chiral order parameter are expected to become critically slowed down. If this is the case the chiral phase transition can occur in a quenched manner and strongly out of equilibrium. Under these circumstances, Wilczek and Rajagopal argued that large domains in which the chiral order parameter could be disoriented with respect to the vacuum could appear[20]. These domains are coherent pion condensates that form after the non-equilibrium phase transition much in the same manner as the correlated domains in condensed matter systems. These pion condensates decay, the neutral pion decays into two photons and the charged pions decay into muons. The pions can then be reconstructed and therefore these disoriented chiral condensates could lead to experimentally observable anomalies in the ratio of the number of neutral to charged

pions. In isospin symmetric condensates the probability for finding a ratio  $R$  of neutral to total (neutral plus charged) is  $P(R) \propto \delta(R - 1/3)$  (for large number of pions) whereas a disoriented chiral condensate leads to  $P(R) \propto 1/\sqrt{R}$ [21].

The possibility of formation of disoriented chiral condensates had been previously conjectured by Bjorken[21] as a potential explanation of CENTAURO events[22], these are cosmic rays events with anomalous neutral to charged pion ratios. This possibility of a distinct signature associated with the chiral phase transition sparked an intense theoretical effort[23, 24]. Several experimental searches are trying to find evidence for this pion condensates or disoriented chiral condensates at CERN-SPS (experiment WA98[25]) at the Tevatron at Fermilab (Minimax experiment[26]), the PHENIX detector at RHIC[27] at BNL can provide an event-by-event analysis of this potential candidates and the ALICE experiment scheduled at CERN-LHC includes the detector CASTOR[28] that will be studying CENTAURO type events.

These disoriented chiral condensates are coherent pion domains and describe the same type of phenomenon of the ferromagnetic domains in quenched ferromagnets or the superfluid domains in He superfluids as described previously. If these condensates are realized during a non-equilibrium stage of the chiral phase transition, they could lead to important experimental probes of this transition and hopefully will be amenable of detection at the RHIC and LHC.

There is an important difference in the dynamics of the chiral phase transition in the Early Universe and at Ultrarelativistic Heavy Ion Colliders. In the Early Universe, the chiral phase transition occurred at a temperature of 150MeV when the Universe was about  $10^{-5}$  seconds old in the radiation dominated era. The size of the Universe at that time was about 10 Km which is much larger than the mean-free path of quarks and gluons  $\approx 10^{-15}$ m and the time scale for cooling  $T/\dot{T} \approx 10^{-5}$ secs is much longer than the relaxation time scale  $\tau_{rel} \approx 10^{-23}$ secs. Therefore in the Early Universe the confining and chiral phase transition occurred in *equilibrium*. These time and length scales must be compared to those in heavy ion collisions: the time scale for cooling from hydrodynamic expansion is few fm/c and the relaxation time scale near phase transitions could be longer and comparable to the lifetime of the quark-gluon plasma, furthermore current numerical estimates determine that the region in which the QGP is formed is about 20 fm. Hence there is a possibility that these phase transitions could be out of equilibrium in heavy ion collisions and that novel phenomena associated with the process of phase ordering and the emergence of pion condensates could be important experimental signatures of the chiral transition.

### 1.3 Early Universe Cosmology:

The COBE satellite mission revolutionized the field of Cosmology. The discovery of temperature fluctuations in the Cosmic Microwave Background (CMB) of  $30\mu K$  imprinted on a blackbody spectrum at  $2.73K$  provides supporting evidence for the main ideas that seek to explain the small inhomogeneities that gave rise to large scale structure formation[3, 4].

Current theoretical ideas maintain that the universe underwent a period accelerated expansion called inflation at an energy scale determined by Grand Unified Theories  $\approx 10^{16}$ Gev. During this period the size of the universe grew by a factor  $e^{60}$  allowing this inflationary scenario to solve the main difficulties of the standard Big Bang Cosmology[3, 4]. Small quantum fluctuations that were present during this epoch of inflation soon became causally disconnected and therefore unaffected by microphysical processes. These fluctuations became in causal contact again at a much later stage of the cosmological evolution, when the Universe was basically dominated by matter. Small fluctuations begin to grow under gravitational instability when they become causally connected again but after the epoch of radiation-matter equality at a temperature of about 10eV and redshift  $z \approx 10^4$ [3]. The COBE observations are sensitive to those fluctuations that have established causal contact again after the epoch of recombination about 300000 years after the Big Bang at a redshift of about  $z \approx 1100$ . Therefore observations of the CMB allow to obtain information on the spectrum of primordial *quantum* fluctuations that were present shortly after the Big Bang.

The first scenario for inflation relied on a supercooled phase transition[3]. Recent detailed studies of the *dynamics* of phase transitions in early universe cosmology[29] allow a reliable calculation of the dynamics including backreaction effects on the metric and a self-consistent evolution

of classical gravity and quantum fields. This approach allows to extract the power spectrum of the primordial perturbations of the metric arising from the quantum fluctuations in the matter fields. These fluctuations are directly related to those of the temperature of the CMB at the scale of recombination and correspond to the Sachs-Wolff plateau in the power spectrum measured by COBE[3, 4]. It is found[29] that the growth of correlated domains after a supercooled phase transition of second order (no metastability) favors a power spectrum with more power on long wavelengths[29] as a consequence of the process of phase ordering. This is a consequence of the instabilities associated with the early stages of the phase ordering dynamics. This 'red' power spectrum is consistent with the results of COBE for the temperature anisotropies, provided that the couplings of the matter field are fine tuned[29].

After inflation the universe underwent several phase transitions. As in all phase transition, fluctuations grow large and may have induced density fluctuations leaving imprints in the CMB as well as acting as seeds of structure formation. How important are these effects on observable quantities is still under investigation [30].

Most works about phase transitions after inflation made the assumption that defects govern such transitions. There is a very important difference between the fluctuations in the inflationary and the defects scenarios. In inflation, the quantum fluctuations become causally disconnected[3, 4] and therefore their evolution is very simple until they become causally connected again because these fluctuations are not influenced by microphysical processes during the period of acausal evolution. Contrary to this dynamics, defects are always causal and are constantly influenced by microphysical processes. Their evolution must be followed dynamically from the time at which a network of defects is formed, at a GUT scale, all the way up to the time scale at which they result in the formation of large scale structure—several billion years later!. Obviously this is an enormous dynamical range, however, detailed computer simulations reveal that a *scaling* solution emerges (for details see[2]) determined by a dynamical length scale. The results of numerical studies suggest that this dynamical length scale is completely determined by the size of the causal horizon at a given time (see the later section on Cosmology for details on causal horizons). However, recent works on defects give a clear indication that current models of defects contradict CMB measurements [31]

The emergence of this length scale through the dynamical process of phase ordering is exactly the same that has been previously discussed within the context of condensed matter systems. Current ground based and balloon borne experiments along with large scale surveys and future satellite missions will provide a flood of data that will support or falsify current theoretical ideas on large scale structure formation and temperature anisotropies. Thus an important theoretical effort goes in providing reliable predictions on the power spectrum of primordial quantum fluctuations. It is a tantalizing possibility that these cosmological observations could provide a definite evidence for cosmological phase transitions.

## 1.4 ...Therefore

We have seen in detail that the dynamics of phase ordering and evolution after non-equilibrium phase transitions are of fundamental importance in a wide range of energies from meV, in Condensed Matter, through GeV in the physics of the Quark Gluon Plasma and the Chiral Phase Transition all the way to GUT's ( $10^{16}$  GeV) in Early Universe Cosmology. An important technical aspect in the study of these phenomena is their *non-perturbative* nature: in a rapid phase transition (of typical second order without metastable states) small amplitude long-wavelength fluctuations become unstable (this will be understood in detail below) and grow in time. The amplitude of these fluctuations must grow until they begin to sample the broken symmetry states of thermodynamic equilibrium.

Having discussed in some detail the importance of the dynamics of symmetry breaking phase transitions out of equilibrium within important settings and their experimental study, we now provide some of the technical aspects that help clarify the phenomena and their quantitative study. We begin by describing a phenomenological approach to phase ordering kinetics in condensed matter systems, highlighting the important ingredients and concepts. We then move on to furnish a quantitative approach to the study of the non-equilibrium dynamics in quantum field theory to

compare some striking similarities to condensed matter and also to contrast some important and relevant differences. The main point for delving into some technical details is to emphasize many *robust* features of the dynamics of symmetry breaking and phase ordering,

- The early stages of phase ordering are determined by linear (spinodal) instabilities. Long-wavelength fluctuations become unstable and grow.
- The emergence of a *dynamical* length scale. This scale represents the average size of the ordered domains and grows in time, eventually at asymptotically long times becoming macroscopically large.
- Associated with this dynamical length scale there is *dynamical scaling*, asymptotically this length scale determines the behavior of correlation functions.
- Coarsening: the growth in time of this correlation length translates in that the peak of the power spectrum moves towards longer wavelength, resulting in a sharp ‘Bragg peak’ at asymptotically long times. This Bragg peak reflects the onset of condensates corresponding to ordered regions of macroscopic size.

As we will see in detail, the phenomenological description in condensed matter systems is **very different** from the microscopic description in quantum field theory. Nevertheless we find that despite these important differences the above features are fairly robust and common to all of the situations studied. Only an excursion into the technical details can reveal in full force these very important and remarkable features.

## 2 Phenomenology of phase ordering dynamics in Condensed Matter:

The phenomenological description of phase ordering kinetics begins with a coarse grained local free energy functional of a (coarse grained) local order parameter  $M(\vec{r})$  [7, 8] which determines the *equilibrium* states. In Ising-like systems this  $M(\vec{r})$  is the local magnetization (averaged over many lattice sites), in binary fluids or alloys it is the local concentration difference, in superconductors is the local gap, in superfluids is the condensate fraction etc. The typical free energy is (phenomenologically) of the Landau-Ginzburg form:

$$\begin{aligned} F[M] &= \int d^d \vec{x} \left\{ \frac{1}{2} [\nabla M(\vec{x})]^2 + V[M(\vec{x})] \right\} \\ V[M] &= \frac{1}{2} r(T) M^2 + \frac{\lambda}{4} M^4 ; \quad r(T) = r_0(T - T_c) \end{aligned} \quad (1)$$

Fig. 1 depicts  $V[M]$  for  $T > T_c$  and  $T < T_c$ . The equilibrium states for  $T < T_c$  correspond to the broken symmetry states with  $M = \pm M_0(T)$  with

$$M_0(T) = \begin{cases} 0 & \text{for } T > T_c \\ \sqrt{\frac{r_0}{\lambda}}(T_c - T)^{\frac{1}{2}} & \text{for } T < T_c \end{cases} \quad (2)$$

Below the critical temperature the potential  $V[M]$  features a non-convex region with  $\partial^2 V[M]/\partial M^2 < 0$  for

$$-M_s(T) < M < M_s(T) ; \quad M_s(T) = \sqrt{\frac{r_0}{3\lambda}}(T - T_c)^{\frac{1}{2}} \quad (T < T_c) \quad (3)$$

this region is called the spinodal region and corresponds to thermodynamically unstable states. The lines  $M_s(T)$  vs.  $T$  and  $M_0(T)$  vs.  $T$  [see eq.(2)] are known as the classical spinodal and coexistence lines respectively. Fig. 2 displays the classical spinodal and coexistence curves for the potential  $V[M]$  in (1).

The states between the spinodal and coexistence lines are metastable (in mean-field theory). As the system is cooled below  $T_c$  into the unstable region inside the spinodal, the *equilibrium* state

of the system is a coexistence of phases separated by domains and the concentration of phases is determined by the Maxwell construction and the lever rule.

**Question:** How to describe the *dynamics* of the phase transition and the process of phase separation?

**Answer:** A phenomenological but experimentally succesful description, Time Dependent Ginzburg-Landau theory (TDGL) where the basic ingredient is Langevin dynamics[7]-[10]

$$\frac{\partial M(\vec{r}, t)}{\partial t} = -\Gamma[\vec{r}, M] \frac{\delta F[M]}{\delta M(\vec{r}, t)} + \eta(\vec{r}, t) \quad (4)$$

with  $\eta(\vec{r}, t)$  a stochastic noise term, which is typically assumed to be white (uncorrelated) and Gaussian and obeying the fluctuation-dissipation theorem:

$$\langle \eta(\vec{r}, t) \eta(\vec{r}', t') \rangle = 2T \Gamma(\vec{r}) \delta^3(\vec{r} - \vec{r}') \delta(t - t') \quad ; \quad \langle \eta(\vec{r}, t) \rangle = 0 \quad (5)$$

the averages  $\langle \dots \rangle$  are over the Gaussian distribution function of the noise. There are two important cases to distinguish: **NCOP**: Non-conserved order parameter, with  $\Gamma = \Gamma_0$  a constant independent of space, time and order parameter, and which can be absorbed in a rescaling of time. **COP**: Conserved order parameter with

$$\Gamma[\vec{r}] = -\Gamma_0 \nabla_{\vec{r}}^2$$

where  $\Gamma_0$  could depend on the order parameter, but here we will restrict the discussion to the case where it is a constant. In this latter case the average over the noise of the Langevin equation can be written as a conservation law

$$\begin{aligned} \frac{\partial M}{\partial t} &= -\nabla \cdot \vec{J} + \eta \Rightarrow \frac{\partial}{\partial t} \langle \int d^3 \vec{r} M(\vec{r}, t) \rangle = 0 \\ \vec{J} &= \vec{\nabla}_{\vec{r}} \left[ -\Gamma_0 \frac{\delta F[M]}{\delta M} \right] \equiv \vec{\nabla}_{\vec{r}} \mu \end{aligned} \quad (6)$$

where  $\mu$  is recognized as the chemical potential. Examples of the NCOP are the magnetization in ferromagnets, the gap in superconductors and the condensate density in superfluids (the **total** particle number is conserved but not the condensate fraction), of the COP: the concentration difference in binary fluids or alloys. For a quench from  $T > T_c$  deep into the low temperature phase  $T \rightarrow 0$  the thermal fluctuations are suppressed after the quench and the noise term is irrelevant. In this situation of experimental relevance of a deep quench the dynamics is now described by a deterministic equation of motion,

for **NCOP**:

$$\frac{\partial M}{\partial t} = -\Gamma_0 \frac{\delta F[M]}{\delta M} \quad (7)$$

for **COP**:

$$\frac{\partial M}{\partial t} = \nabla^2 \left[ \Gamma_0 \frac{\delta F[M]}{\delta M} \right] \quad (8)$$

which is known as the Cahn-Hilliard equation[7, 8]. In both cases the equations of motion are purely diffusive

$$\frac{dF}{dt} = \int d^3 r \frac{\delta F[M]}{\delta M(\vec{r}, t)} \frac{\partial M(\vec{r}, t)}{\partial t} = -\Gamma_0 \left\{ \begin{array}{l} \int d^3 r \left( \frac{\delta F}{\delta M} \right)^2 \text{ NCOP} \\ \int d^3 r \left( \vec{\nabla} \frac{\delta F}{\delta M} \right)^2 \text{ COP} \end{array} \right. \quad (9)$$

and in both cases  $\frac{dF}{dt} < 0$ . Thus, the energy is always diminishing and there is no possibility of increasing the free energy. Thus overbarrier thermal activation cannot be described in the absence of thermal noise, which is clear since thermal activation is mediated by large thermal fluctuations. The fact that this phenomenological description is purely dissipative with an ever diminishing free energy is one of the fundamental differences with the quantum field theory description studied in the next sections.

## 2.1 Critical slowing down in NCOP:

Critical slowing down of long-wavelength fluctuations is built in the TDGL description. Consider the case of NCOP and linearize the TDGL equation above the critical temperature for small amplitude fluctuations near  $M = 0$ . Neglecting the noise term for the moment and taking the Fourier transform of the small amplitude fluctuations we find

$$\frac{dm_k(t)}{dt} \approx -\Gamma_0 [k^2 + r_0(T - T_c)] m_k(t) \quad (10)$$

showing that long-wavelength small amplitude fluctuations relax to equilibrium  $m_k = 0$  on a time scale given by

$$\tau_k \propto [k^2 + r_0(T - T_c)]^{-1} \quad (11)$$

As  $T \rightarrow T_c^+$  the long-wavelength modes are critically slowed down and relax to equilibrium on very long time scales. Therefore a TDGL description leads to the conclusion that if the cooling rate is finite, the long-wavelength modes will fall out of LTE and become quenched. As the temperature falls below the critical, these modes will become unstable and will grow exponentially.

## 2.2 Linear instability analysis:

Let us consider now the situation for  $T \ll T_c$  and neglect the thermal noise. The early time evolution after the quench is obtained by linearizing the TDGL equation around a homogeneous mean field solution  $M_o(t)$ . Writing

$$M(\vec{r}, t) = M_o(t) + \frac{1}{\sqrt{\Omega}} \sum_{\vec{k} \neq 0} m_k(t) e^{i\vec{k} \cdot \vec{r}} \quad (12)$$

where  $\Omega$  is the volume of the system, and considering only the linear term in the fluctuations  $m_k(t)$  the linearized dynamics is the following: **COP**: for  $M_o(t)$  the conservation gives

$$\frac{dM_o(t)}{dt} = 0$$

since  $M_o$  is the volume integral of the order parameter [see eq.(6)] and for the fluctuations we obtain

$$\frac{dm_k(t)}{dt} = \omega(k) m_k(t) \quad ; \quad \omega(k) = -\Gamma_0 k^2 \left[ k^2 + \frac{\partial^2 V[M]}{\partial M^2} \Big|_{M_o} \right] \quad (13)$$

In the spinodal region  $\frac{\partial^2 V[M]}{\partial M^2} \Big|_{M_o} < 0$  there is a band of unstable wave vectors  $k^2 < \left| \frac{\partial^2 V[M]}{\partial M^2} \Big|_{M_o} \right|$  for which the frequencies are positive and the fluctuations away from the mean field grow exponentially.

**NCOP**: separate the  $\vec{k} \neq 0$  from the  $\vec{k} = 0$  in the linearized equation of motion:

$$\frac{dM_o(t)}{dt} = -\Gamma_0 \frac{dV[M]}{dM} \Big|_{M_o} \quad (14)$$

$$\frac{dm_k(t)}{dt} = -\Gamma_0 \left[ \frac{\delta F[M]}{\delta M} \right]_{M_o(t)} m_k(t) = -\Gamma_0 \left[ k^2 + \frac{\partial^2 V[M]}{\partial M^2} \Big|_{M_o} \right] \quad (15)$$

whereas the first equation (14) determines that  $M_o(t)$  rolls down the potential hill towards the equilibrium solution, the second equation also displays the linear instabilities for the same band of wave vectors as in the COP in the spinodal region  $|M_o(t)| \leq M_s(T)$  [see eq. (3)] for which the fluctuations grow exponentially in time. Thus in the linearized approximation for both NCOP and the COP the spinodal instabilities are manifest as exponentially growing fluctuations. These instabilities are the hallmark of the process of phase separation and are the early time indications of the formation and growth of correlated regions which will be understood in an exactly solvable example below.

### 2.3 The scaling hypothesis: dynamical length scales for ordering

The process of ordering is described by the system developing ordered regions or domains that are separated by walls or other type of defects. The experimental probe to study the domain structure and the emergence of long range correlations is the equal time pair correlation function

$$C(\vec{r}, t) = \langle M(\vec{r}, t) M(\vec{0}, t) \rangle \quad (16)$$

where  $\langle \dots \rangle$  stands for the statistical ensemble average in the initial state (or average over the noise in the initial state before the quench) and will become clear(er) below. It is convenient to expand the order parameter in Fourier components

$$M(\vec{r}, t) = \frac{1}{\sqrt{\Omega}} \sum_{\vec{k}} m_{\vec{k}}(t) e^{i\vec{k} \cdot \vec{r}}$$

and to consider the spatial Fourier transform of the pair correlation function

$$S(\vec{k}, t) = \langle m_{\vec{k}}(t) m_{-\vec{k}}(t) \rangle \quad (17)$$

known as the **structure factor** or power spectrum which is experimentally measured by neutron (in ferromagnets) or light scattering (in binary fluids)[12]. The scaling hypothesis introduces a dynamical length scale  $L(t)$  that describes the typical scale of a correlated region and proposes that

$$C(\vec{r}, t) = f\left(\frac{|\vec{r}|}{L(t)}\right) \Rightarrow S(\vec{k}, t) = L^d(t) g(kL(t)) \quad (18)$$

where  $d$  is the spatial dimensionality and  $f$  and  $g$  are scaling functions. Ultimately scaling is confirmed by experiments and numerical simulations and theoretically it emerges from a renormalization group approach to dynamical critical phenomena which provides a calculational framework to extract the scaling functions and the deviations from scaling behavior[7]. This scaling hypothesis describes the process of phase ordering as the formation of ordered ‘domains’ or correlated regions of typical spatial size  $L(t)$ . For NCOP typical growth laws are  $L(t) \approx t^{1/2}$  (with some systems showing weak logarithmic corrections) and  $L(t) \approx t^{1/3}$  for scalar and  $\approx t^{1/4}$  for vector order parameter in the COP case[7, 9, 10].

### 2.4 An exactly solvable (and relevant) example: the Large $N$ limit

We consider the case where the order parameter has  $N$ -components and transforms as a vector under rotations in an  $N$ -dimensional Euclidean space, i.e.  $\vec{M}(\vec{r}, t) = (M_1(\vec{r}, t), M_2(\vec{r}, t), \dots, M_N(\vec{r}, t))$ . For  $N = 1$  an example is the Ising model, for  $N = 2$  superfluids or superconductors (where the components are the real and imaginary part of the condensate fraction or the complex gap respectively),  $N = 3$  is the spin one Heisenberg antiferromagnet, etc. For  $N = 1$  the topological defects are domain walls (topological in one spatial dimension), for  $N = 2$  they are vortices in  $d = 2$  and vortex lines in  $d = 3$ , for  $N = d = 3$  the topological defects are monopoles or skyrmions which are possible excitations in Quantum Hall systems and also appear in nematic liquid crystals[2]. For  $N \rightarrow \infty$  and fixed  $d$  no topological defects exist. However the exact solution of the large  $N$  model gives insight and is in fairly good agreement with growth laws for fixed  $N$  systems which had been studied experimentally and numerically[7, 10]. In cosmological space-times it has been implemented to study the collapse of texture-like configurations[1, 32, 33] (see later). In quantum field theory the non-equilibrium dynamics of phase transitions has been studied in Minkowsky and cosmological space-times[24, 34, 35, 36, 37, 38]. The large  $N$  limit is an exactly solvable limit that serves as a testing ground for establishing the fundamental concepts and that can be systematically improved in a consistent  $1/N$  expansion. It provides a consistent formulation which is *non-perturbative*, renormalizable and numerically implementable and has recently been invoked in novel studies of non-equilibrium dynamics in quantum spin glasses and disordered systems[39].

The exact solution for the dynamics in the large  $N$  limit, being available both in the condensed matter TDGL description of phase ordering kinetics and in Quantum Field Theory in Minkowsky

and Cosmological space times, allow us to compare *directly* the physics of phase ordering in these situations. Thus we begin by implementing this scheme in the NCOP case for the TDGL description.

What is the  $\langle \dots \rangle$  in the equations of the previous section?: consider that *before* the quench the system is in equilibrium in the *disordered* phase at  $T \gg T_c$  and with a very short correlation length ( $\xi(T) \approx 1/T$ ). The ensemble average in this initial state is therefore

$$\begin{aligned} \langle M^i(\vec{r}, 0) M^j(\vec{r}', 0) \rangle &= \Delta \delta^{ij} \delta^3(\vec{r} - \vec{r}') \\ \langle M^i(\vec{r}, 0) \rangle &= 0 \end{aligned} \quad (19)$$

where  $\Delta$  specifies the initial correlation. Now consider a critical quench where the system is rapidly cooled through the phase transition to almost zero temperature but in the *absence* of explicit symmetry breaking fields (for example a magnetic field). The average of the order parameter will remain zero through the process of spinodal decomposition and phase ordering. During the initial stages, linear instabilities will grow exponentially with  $m_k^i(t) \approx m_k^i(0) e^{\omega(k)t}$ ;  $\omega(k) = k^2 - r(0)$  for  $k^2 < r(0)$  and at early times

$$\langle m_{\vec{k}}^i(t) m_{-\vec{k}}^j(t) \rangle \approx \Delta e^{2\omega(k)t} \quad (20)$$

hence fluctuations begin to grow exponentially and eventually will sample the broken symmetry states and the exponential growth must shut-off. The large  $N$  limit is implemented by writing the potential term in the free energy as

$$V[\vec{M}] = -\frac{r(T)}{2} \vec{M}^2 + \frac{\lambda}{4N} (\vec{M}^2)^2; \quad \vec{M}^2 = \vec{M} \cdot \vec{M} \quad (21)$$

where  $\lambda$  is kept finite in the large  $N$  limit. We will focus on the NCOP case with a quench to zero temperature and rescale the order parameter, time and space as

$$\vec{M} = \sqrt{\frac{r(0)}{\lambda}} \vec{\eta}; \quad r(0) \Gamma_0 t = \tau; \quad \sqrt{r(0)} \vec{x} = \vec{z} \quad (22)$$

after which the evolution equation for the NCOP case becomes

$$\frac{\partial \vec{\eta}}{\partial \tau} = \nabla^2 \vec{\eta} + \left(1 - \frac{\vec{\eta}^2}{N}\right) \vec{\eta} \quad (23)$$

where derivatives are now with respect to the rescaled variables. The large  $N$  limit is solved by implementing a Hartree-like factorization[7]

$$\vec{\eta}^2 \rightarrow \langle \vec{\eta}^2 \rangle = N \langle \eta_i^2 \rangle \quad \text{no sum over } i \quad (24)$$

Then for each component the NCOP equation becomes

$$\frac{\partial \eta_i}{\partial \tau} = [\nabla^2 + M^2(t)] \eta_i \quad (25)$$

$$M^2(t) = 1 - \langle \eta_i^2 \rangle \quad (26)$$

the eq.(26) is a *self-consistent* condition that must be solved simultaneously with the equation of motion for the components. Thus the large  $N$  approximation linearizes the problem at the expense of a self-consistent condition. The solution for each component is obviously

$$\eta_{i,\vec{k}}(\tau) = \eta_{i,\vec{k}}(0) e^{-k^2\tau + b(\tau)}; \quad b(\tau) = \int_0^\tau M^2(\tau') d\tau' \quad (27)$$

Consider for a moment that the  $\vec{k} = 0$  mode is slightly displaced at the initial time, then it will roll down the potential hill to a final equilibrium position for which  $M^2(\infty) \eta_i(\infty) = 0$  (so the time



derivative vanishes in equilibrium). If  $\eta_i(\infty) \neq 0$  is a broken symmetry minimum of the free energy, then  $M^2(\tau) \rightarrow 0$  when  $\tau \rightarrow \infty$ . This is the statement of Goldstone's theorem that guarantees that the perpendicular fluctuations are soft modes. This asymptotic limit allows the solution of the self-consistent condition

$$M^2(\tau) = 1 - \langle \eta_i^2(\tau) \rangle = 1 - \Delta e^{2b(\tau)} \int \frac{d^d k}{(2\pi)^d} e^{-k^2 \tau} = 1 - \Delta e^{2b(\tau)} (8\pi t)^{-\frac{d}{2}} \quad (28)$$

The vanishing of the right hand side in the asymptotic time regime leads to the self-consistent solution

$$b(\tau) \rightarrow \frac{d}{4} \ln \left[ \frac{\tau}{\tau_0} \right] \Rightarrow M^2(\tau) \rightarrow \frac{d}{4\tau} \quad (29)$$

where  $\tau_0$  is a constant related to  $\Delta$ . This self-consistent solution results in the following asymptotic behavior

$$\eta_{i,\vec{k}}(\tau) \rightarrow \eta_{i,\vec{k}}(0) \left( \frac{\tau}{\tau_0} \right)^{\frac{d}{4}} e^{-k^2 \tau} \quad (30)$$

Introducing the *dynamical length scale*  $L(\tau) = \tau^{\frac{1}{2}}$  it is straightforward to find the structure factor and the pair correlation function

$$S(\vec{k}, t) \propto L^d(t) e^{-2(kL(t))^2} \quad (31)$$

$$C(\vec{r}, t) \propto e^{-\frac{r^2}{8L^2(t)}} ; \quad L(t) = t^{\frac{1}{2}} \quad (32)$$

This behavior *should not* be interpreted as diffusion, because of the  $L^d(t)$  in eqn. (31) which is a result of the self-consistent condition.

#### Important Features:

- The 'effective squared mass'  $M^2(t) \xrightarrow{t \rightarrow \infty} 0$ : asymptotically there are massless excitations identified as Goldstone bosons.
- Since  $M^2(t) \rightarrow 0$  asymptotically, the self-consistent condition results in that  $\langle \tilde{M}^2 \rangle \rightarrow Nr(0)/\sqrt{\lambda}$ , i.e. the fluctuations sample the broken symmetry states, which are equilibrium minima of the free energy. These fluctuations begin to grow exponentially at early times due to spinodal instabilities.
- A dynamical correlation length emerges  $L(t) = t^{1/2}$  which determines the size of the correlated regions or 'domains'. A scaling solution emerges asymptotically with the natural scale determined by the size of the ordered regions. These regions grow with this law until they become macroscopically large. Although this a result obtained in the large  $N$  limit, similar growth laws had been found for NCOP both analytically and numerically for  $N = 1$  etc.[7]
- **Coarsening:** The expression for the structure factor (31) shows that at large times only the very small wavevectors contribute to  $S(\vec{k}, t)$ , however the self-consistency condition forces the  $\int k^{d-1} dk S(k, t) \rightarrow \text{constant}$  thus asymptotically  $k^{d-1} S(k, t)$  is peaked at wavevectors  $k \approx L^{-1}(t)$  with an amplitude  $L^d(t)$  thus becoming a *delta function*  $S(\vec{k}, t) \xrightarrow{t \rightarrow \infty} \delta^d(\vec{k})$ . The position of the peak in  $S(\vec{k}, t)$  moving towards longer wavelength is the phenomenon of coarsening and is observed via light scattering. At long times a zero momentum condensate is formed[10] and a Bragg peak develops at zero momentum, this condensate however grows as a power of time and only becomes macroscopic at asymptotically large times. Coarsening is one of the experimental hallmarks of the process of phase ordering, revealed for example in light scattering[12] and is found numerically in many systems[7]. Thus the large  $N$  limit, although not being able to describe topological defects offers a very good description of the ordering dynamics.

### 3 Phase ordering in Quantum Field Theory I: Minkowski space-time

#### 3.1 A quench in Q.F.T.

Although the phenomenological Time Dependent Landau Ginzburg theory is a succesful description of phase ordering kinetics in condensed matter systems, there is no first principle derivation from a microscopic theory of these equations of motion. Whereas microscopic descriptions either based on classical or quantum Hamiltonians lead to time reversal invariant equations of motion, the TDGL equations are first order in the time derivative and therefore purely dissipative.

A first principles, microscopic description of a quantum theory must begin with the Heisenberg equations of motion for operators or the Schroedinger or quantum Liouville equations for the quantum states or density matrix that describes the system. In this section we provide an introduction to the treatment of strongly out of equilibrium situations, in particular that of a ‘quench’ in a quantum field theory system.

This is the situation studied in[40] for the dynamics of formation and evolution of disoriented chiral condensates during the chiral phase transition.

The dynamics is completely determined by the microscopic field theoretical Hamiltonian. For a simple scalar theory the Hamiltonian operator is given by

$$\hat{H} = \int d^3x \left\{ \frac{1}{2} \Pi^2(\vec{x}, t) + \frac{1}{2} [\vec{\nabla} \Phi(\vec{x}, t)]^2 + V[\Phi(\vec{x}, t)] \right\} \quad (33)$$

where  $\Phi$  is the quantum mechanical field and  $\Pi$  its canonical momentum. We want to describe a quenched scenario where the initial state of the system for  $t < 0$  is the ground state (or density matrix, see later) of a Hamiltonian for which the potential is convex for all values of the field, for example that of an harmonic oscillator, in which case the wave function(al)  $\Psi[\Phi]$  is a Gaussian centered at the origin. At  $t = 0$  the potential is changed so that for  $t > 0$  it allows for broken symmetry states. This can be achieved for example by the following form

$$V[\Phi] = \frac{1}{2} m^2(t) \Phi^2 + \frac{\lambda}{4} \Phi^4 \quad (34)$$

$$m^2(t) = \begin{cases} +m_0^2 > 0 & \text{for } t < 0 \\ -m_0^2 < 0 & \text{for } t > 0 \end{cases} \quad (35)$$

thus the potential in Fig. 1 changes *suddenly* from  $T > T_c$  to  $T < T_c$ . Although in Minkowski space-time this is an *ad-hoc* choice of a time dependent potential that mimics the quench[41], we will see in the next section that in a cosmological setting the mass term naturally depends on time through the temperature dependence and that it changes sign below the critical temperature as the Universe cools off. Most of the results obtained in Minkowski space-time will translate onto analogous results in a Friedmann-Robertson-Walker cosmology. Unlike the phenomenological (but succesful) description of the dynamics in condensed matter systems, in a microscopic quantum theory the dynamics is completely determined by the Schrödinger equation for the time evolution of the wave function or alternatively the Liouville equation for the evolution of the density matrix in the case of mixed states. We will cast our study in terms of a density matrix in general, such a density matrix could describe pure or mixed states and obeys the quantum Liouville equation

$$i \frac{\partial \hat{\rho}(t)}{\partial t} = [\hat{H}(t), \hat{\rho}(t)] \quad (36)$$

**Question:** How does the wave function(al) or the density matrix evolve after a quench?

#### 3.2 A simple quantum mechanical picture:

In order to gain insight into the above question, let us consider a simple case of one quantum mechanical degree of freedom  $q$  and the quench is described in terms of an harmonic oscillator

with a time dependent frequency  $\omega^2(t) = -\epsilon(t) \omega_0^2$ ;  $\omega_0^2 > 0$  with  $\epsilon(t)$  the sign function, so that  $\omega^2(t < 0) > 0$ ;  $\omega^2(t > 0) < 0$ . Furthermore let us focus on the evolution of a pure state (the density matrix is simple the product of the wave function and its complex conjugate). Consider that at  $t < 0$  the wave function corresponds to the ground state of the (upright) harmonic oscillator. For  $t > 0$  the wave function obeys

$$i \frac{\partial \Psi[q, t]}{\partial t} = \left[ -\frac{1}{2} \frac{d^2}{dq^2} - \frac{1}{2} \omega_0^2 q^2 \right] \Psi[q, t] \quad (37)$$

Since the initial wave function is a gaussian and under time evolution with a quadratic Hamiltonian Gaussians remain Gaussians, the solution of this Schrödinger equation is given by

$$\Psi[q, t] = N(t) e^{-\frac{A(t)}{2} q^2} \quad (38)$$

$$\frac{d \ln N(t)}{dt} = -\frac{i}{2} A(t) \quad (39)$$

$$i \frac{dA}{dt} = A^2 + \omega_0^2$$

Separating the real and imaginary parts of  $A(t)$  it is straightforward to find that  $|N(t)|^4 / \text{Re}[A(t)]$  is constant, a consequence of unitary time evolution. Eq.(40) can be cast in a more familiar form by a simple substitution

$$A(t) = -i \frac{\dot{\phi}(t)}{\phi(t)} \Rightarrow \ddot{\phi}(t) - \omega_0^2 \phi(t) = 0 \quad (40)$$

where the equation for  $\phi$  was obtained by inserting the above expression for  $A(t)$  in (40). The solution is  $\phi(t) = a e^{\omega_0 t} + b e^{-\omega_0 t}$  featuring exponential growth. This is the quantum mechanical analog of the spinodal instabilities described in the previous section. The equal time two-point function is given by

$$\langle q^2 \rangle(t) = A_R^{-1}(t) = |\phi(t)|^2 \approx e^{2\omega_0 t} \quad (41)$$

The width of the Gaussian state increases in time (while the amplitude decreases to maintain a constant norm) and the quantum fluctuations grow exponentially. As the Gaussian wave function spreads out the probability for finding configurations with large amplitude of the coordinates increases. These is the quantum mechanical translation of the linear spinodal instabilities. When the non-linear contributions to the quantum mechanical potential are included the single particle quantum mechanical wave function will simply develop two peaks and eventually re-collapse by focusing near the origin undergoing oscillatory motion between ‘collapses’ and ‘revivals’. In the case of a full quantum field theory there are infinitely many degrees of freedom and the energy is transferred between many modes. This simple quantum mechanical example paves the way for understanding in a simple manner the main features of a quench in the large  $N$  limit in quantum field theory, to which we now turn our attention.

### 3.3 Back to the original question: Large $N$ in Q.F.T.

We now consider the large  $N$  limit of a full Q.F.T. in which

$$\vec{\Phi}(\vec{x}, t) = (\Phi_1(\vec{x}, t), \Phi_2(\vec{x}, t), \dots, \Phi_N(\vec{x}, t)) \quad (42)$$

and similarly for the canonical momenta  $\vec{\Pi}$ . The Hamiltonian operator is of the form (33) with

$$V[\vec{\Phi}] = \frac{1}{2} m^2(t) \vec{\Phi} \cdot \vec{\Phi} + \frac{\lambda}{8N} [\vec{\Phi} \cdot \vec{\Phi}]^2 \quad (43)$$

with  $m^2(t)$  given by (35). Let us focus on the case in which the initial state pure and symmetric, i.e.  $\langle \Phi \rangle = 0$ , with  $\langle \dots \rangle$  being the expectation value in this initial state. The more complicated case of a mixed state, described by a density matrix is studied in detail in [35, 36, 37] and the main

features are the same as those revealed by the simpler scenario of a pure state. The large  $N$  limit is implemented in a similar manner as in the TDGL example, via a Hartree like factorization

$$(\vec{\Phi} \cdot \vec{\Phi})^2 \rightarrow 2 \langle \vec{\Phi} \cdot \vec{\Phi} \rangle \vec{\Phi} \cdot \vec{\Phi} \quad (44)$$

where the expectation value is in the time evolved quantum state (in the Schrödinger picture) or in the initial state of the Heisenberg operators (in the Heisenberg picture). Via this factorization the Hamiltonian becomes quadratic at the expense of a self-consistent condition as it will be seen below. It is convenient to introduce the spatial Fourier transform of the fields as

$$\vec{\Phi}(\vec{x}, t) = \frac{1}{\sqrt{\Omega}} \sum_{\vec{k}} \vec{\Phi}_{\vec{k}}(t) e^{i\vec{k} \cdot \vec{x}} \quad (45)$$

with  $\Omega$  the spatial volume, and a similar expansion for the canonical momentum  $\Pi(\vec{x}, t)$ . The Hamiltonian becomes

$$H = \sum_{\vec{k}} \left\{ \frac{1}{2} \vec{\Pi}_{\vec{k}} \cdot \vec{\Pi}_{-\vec{k}} + \frac{1}{2} W_k^2(t) \vec{\Phi}_{\vec{k}} \cdot \vec{\Phi}_{-\vec{k}} \right\} \quad (46)$$

$$W_k^2(t) = m^2(t) + k^2 + \frac{\lambda}{2N} \int \frac{d^3 k}{(2\pi)^3} \langle \vec{\Phi}_{\vec{k}} \cdot \vec{\Phi}_{-\vec{k}} \rangle(t) \quad (47)$$

The problem now has decoupled in a set of infinitely many harmonic oscillators, that are only coupled through the self-consistent condition in the frequencies (47). To induce a quench, the time dependent mass term has the form proposed in eq. (35).

Just as in the simple quantum mechanical case, we consider the initial state to be a Gaussian centered at the origin in field space, which is the ground state of the (upright) harmonic oscillators for  $t < 0$ . Since a Gaussian is always a Gaussian under time evolution with a quadratic Hamiltonian, we propose the wave function(al) that describes the (pure) quantum mechanical state to be given by

$$\Psi[\vec{\Phi}, t] = \Pi_k \left\{ N_k(t) e^{-\frac{A_k(t)}{2} \vec{\Phi}_{\vec{k}} \cdot \vec{\Phi}_{-\vec{k}}} \right\} ; \quad A_k(t=0) = W_k(t < 0) \quad (48)$$

Time evolution of this wavefunction(al) is determined by the Schrödinger equation: in the Schrödinger representation the canonical momentum becomes a differential (functional) operator,  $\vec{\Pi}_{\vec{k}} \rightarrow -i\delta/\delta\vec{\Phi}_{-\vec{k}}$  and the Schrödinger equation becomes a functional differential equation. Comparing the powers of  $\vec{\Phi}_{\vec{k}}$  in this differential equation, one obtains the following evolution equations for  $N_k(t)$  and  $A_k(t)$

$$\frac{d}{dt} \ln N_k(t) = -\frac{i}{2} A_k(t) \quad (49)$$

$$i \frac{dA_k(t)}{dt} = A_k^2(t) - W_k^2(t) \quad (50)$$

As in the single particle case, the constancy of  $|N_k(t)|^4 / \text{Re}[A_k(t)]$  is a consequence of unitary time evolution. The non-linear equation for the kernel  $A_k(t)$  can be simplified just as in the single particle case by writing

$$A_k(t) = -i \frac{\dot{\phi}_k(t)}{\phi_k(t)} \Rightarrow \ddot{\phi}_k(t) + W_k^2(t) \phi_k(t) = 0 \quad (51)$$

and taking the expectation value of  $\Phi^2$  in this state we obtain

$$\langle \vec{\Phi}_{\vec{k}} \cdot \vec{\Phi}_{-\vec{k}} \rangle(t) = N |\phi_k(t)|^2 \quad (52)$$

Hence we find a self-consistent condition much like the one obtained in the large  $N$  limit for TDGL. The equations for the mode functions and the self-consistent condition for  $t > 0$  are therefore given by

$$\ddot{\phi}_k(t) + [k^2 + M^2(t)] \phi_k(t) = 0 \quad (53)$$

$$M^2(t) = -m_0^2 + \frac{\lambda}{2} \int \frac{d^3 k}{(2\pi)^3} |\phi_k(t)|^2 \quad (54)$$

where the integral in the self-consistent term in (54) is simply  $\langle \Phi_i^2 \rangle$ . There are two fundamental *differences* between the quantum dynamics determined by the equations of motion and the classical dissipative dynamics of the TDGL phenomenological description given in sec. II:

- The equations of motion and the self-consistency condition equations (53)-(54) lead immediately to the conservation of energy[34, 35].
- The evolution equations are *time reversal invariant*.

These properties must be contrasted to the purely dissipative evolution dictated by the TDGL equations as is clear from eq. (9). Consider a very weakly coupled theory  $\lambda \ll 1$  and very early times, then the self-consistent term can be neglected and we see that for  $k^2 < m_0^2$  the modes grow exponentially. This instability again is the manifestation of spinodal growth[42, 43, 44, 35, 36]. Since the mode functions grow exponentially, fairly soon, at a time scale  $t_s \approx m_0^{-1} \ln(1/\lambda)$  the self-consistent term begins to cancel the negative mass squared and  $M^2(t)$  becomes smaller. We find numerically that this effective mass vanishes asymptotically, as shown in Fig. 3.

### 3.4 Emergence of condensates and classicality:

The physical mechanism here is similar to that in the classical TDGL, but in terms of quantum fluctuations. The quantum fluctuations with wave vectors inside the spinodally unstable band grow exponentially, these make the  $\langle \Phi^2 \rangle$  self-consistent field to grow non-perturbatively large until when  $\langle \Phi^2 \rangle \approx m_0^2/\lambda$  when the self-consistent (mean) field begins to be of the same order as  $m_0^2$  (the tree level mass term). At this point the *quantum* fluctuations become non-perturbatively large and sample field configurations near the equilibrium minima of the potential. The spinodal instabilities are shutting off since the effective squared mass  $M^2(t)$  is vanishing.

When  $M^2(t)$  vanishes, the equations for the mode functions become those of a free massless field, with solutions of the form  $\phi_k(t) = A_k e^{ikt} + B_k e^{-ikt}$ , whereas for the  $k = 0$  mode the solution must be of the form  $\phi_0(t) = a + bt$  with  $a, b \neq 0$  since the Wronskian of the mode function and its complex conjugate is a constant. This in turn determines that the low  $k$  (long wavelength) behavior of the mode functions is given by

$$\phi_k(t) = a \cos kt + b \frac{\sin kt}{k} \quad (55)$$

This behavior at long wavelength has a remarkable consequence: at very long time the power spectrum  $|\phi_k(t)|^2$ , which is the equivalent of  $S(k, t)$  for TDGL (see eq. (17)) is dominated by the small  $k$ -region, in particular  $k \ll 1/t$ , with an amplitude that grows quadratically with time. Then the structure factor  $S(\vec{k}, t) = |\phi_k(t)|^2$  features a peak that moves towards longer wavelengths at longer times and whose amplitude grows with time in such a way that asymptotically  $\int_0^\infty k^2 S(\vec{k}, t) dk / 2\pi^2 \rightarrow m_0^2/\lambda$  and the integral is dominated by a very small region in  $k$  that gets narrower at longer times. This is the equivalent of *coarsening* in the TDGL solution in the large  $N$  limit, where the asymptotic time regime was dominated by the formation of a long-wavelength condensate. Fig. 4 shows the power spectrum at two (large) times displaying clearly the phenomenon of coarsening and the formation of a non-perturbative condensate.

The pair correlation function can now be calculated using this power spectrum[36]

$$C(\vec{r}, t) = \frac{1}{2\pi^2 r} \int_0^\infty k \sin kr |\phi_k^2(t)| dk. \quad (56)$$

At long times and distances the integral is dominated by the very long wavelength modes, in particular by the term  $\propto \sin[kt]/k$  of  $\phi_k(t)$ , hence the integral can be done analytically and we find

$$C(\vec{r}, t) = \frac{A}{r} \Theta(2t - r) \quad (57)$$

with  $A$  a constant. This is a remarkable result: the correlation falls off as  $1/r$  inside domains that grow at the speed of light. This correlation function is shown in Fig. 5 at several different (large)

times. This correlation function is of the *scaling form*: introducing the dynamical length scale  $L(t) = t$  it is clear that[36]

$$C(\vec{r}, t) \propto L^{-1}(t) f(r/L(t)) ; \quad f(s) = \frac{\Theta(2-s)}{s} \quad (58)$$

We interpret these ‘domains’ as being a non-perturbative condensate of Goldstone bosons, with a non-perturbatively large number of them  $\propto 1/\lambda$ , such that the mean square root fluctuation of the field samples the (non-perturbative) equilibrium minima of the potential. In particular an important conclusion of this analysis is that the long-wavelength modes acquire very large amplitudes, their phases vary slowly as a function of time (for  $k \ll 1/t$ ), therefore these fluctuations which began their evolution as being quantum mechanical, now have become *classical*.

### 3.5 Coherent Structures

At this point our analysis begs this question. To understand the answer it is convenient to back track the analysis to the beginning. The initial quantum state is given by a the wave-function(al) (48), thus the most probable field configurations found in this ensemble are those whose spatial Fourier transform are given by

$$|\Phi_k| \propto \frac{1}{\sqrt{W_k(t < 0)}} \propto \frac{1}{\sqrt{k^2 + m_0^2}} \quad (59)$$

(restoring  $\hbar$  would multiply  $\Phi_k$  by  $\sqrt{\hbar}$ ). Then typical long-wavelength field configurations that are represented in the quantum ensemble described by this initial wave-function(al) are of rather small amplitude. The initial correlations are also rather short ranged on scales  $m_0^{-1}$ . Under time evolution the probability distribution is given by

$$\mathcal{P}[\Phi, t] = |\Psi[\Phi, t]|^2 = \prod_{i=1}^N \Pi_k \left\{ |N_k(t)|^2 e^{-\frac{|\Phi_k^i(t)|^2}{|\phi_k(t)|^2}} \right\} \quad (60)$$

At times longer than the regime dominated by the exponential growth of the spinodally unstable modes, the power spectrum  $|\phi_k(t)|^2$  obtains the largest support for long wavelengths  $k \ll m_0^2$  and with amplitudes  $\approx m_0^2/\lambda$ . Therefore field configurations with typical spatial Fourier transform  $\phi_k(t)$  are very likely to be found in the ensemble. These field configurations are primarily made of long-wavelength modes and their amplitudes are non-perturbatively large, of the order of the amplitude of the fields in the broken symmetry minima. A typical such configuration can be written as

$$\Phi^i(\vec{x}, t)_{\text{typical}} \approx \sum_k |\phi_k(t)| \cos[\vec{k} \cdot \vec{x} + \delta_k^i] \quad (61)$$

where the phases  $\delta_k^i$  are randomly distributed with a Gaussian probability distribution since the density matrix is gaussian in this approximation. We note that a particular choice of these phases leads to a realization of a likely configuration in the ensemble that *breaks translational invariance*. In fact translations can be absorbed by a change in the phases, thus averaging over these random phases restores translational invariance. Since the quantum state (or density matrix) is translational invariant a particular spatial profile for a field configuration corresponds to a particular representative of the ensemble. Combining all of the above results together we can present the following consistent interpretation of the ordering process and the formation of coherent non-perturbative structures during the dynamics of symmetry breaking in the large  $N$  limit[36] :

- The early time evolution occurs via the exponential growth of spinodally unstable long wavelength modes. This unstable growth leads to a rapid growth of fluctuations  $\langle \Phi^2 \rangle(t)$  which in turn increases the self-consistent contribution and tends to cancel the negative mass squared. The effective mass of the excitations  $-m_0^2 + \frac{\lambda}{2N} \langle \Phi^2 \rangle(t) \rightarrow 0$  and the asymptotic excitations are Goldstone bosons.

- At times larger than the spinodal time  $t_s \approx m_0^{-1} \ln(1/\lambda)$ , the effective mass vanishes and the power spectrum or structure factor  $S(k, t) = |\phi_k(t)|^2$  displays the features of coarsening: a peak that moves towards longer wavelengths and increases in amplitude, resulting in a long-wavelength condensate at asymptotically long times.
- For large time a dynamical correlation length emerges  $L(t) = t$  and at long distances the pair correlation function is of the scaling form  $C(\vec{r}, t) \propto L^{-1}(t) f(r/L(t))$ . The length scale  $L(t)$  determines the size of the correlated regions and determines that these regions grow at the speed of light. Inside these regions there is a non-perturbative condensate of Goldstone bosons with a typical amplitude of the order of the value of the homogeneous field at the equilibrium broken symmetry minima.

The similarity between these results and those of the more phenomenological TDGL description in condensed matter systems is rather striking. The features that are determined by the structure of the quantum field theory are[36]: i) the scaling variable  $s = r/t$  with equal powers of distance and time is a consequence of the Lorentz invariance of the underlying theory, ii) the fact that the pair correlation function vanishes for  $r > 2t$  is manifestly a consequence of causality. An analysis of the correlations and defect density during the spinodal time scale has been performed in[45] and related recent studies had been performed in[46].

## 4 Phase ordering in Quantum Field Theory II: FRW Cosmology

### 4.1 Cosmology 101 (the basics):

On large scales  $> 100$  Mpc the Universe appears to be homogeneous and isotropic as revealed by the isotropy and homogeneity of the cosmic microwave background and some of the recent large scale surveys[1]. The cosmological principle leads to a simple form of the metric of space time, the Friedmann-Robertson-Walker (FRW) metric in terms of a scale factor that determines the Hubble flow and the curvature of spatial sections. Observations seem to favor a flat Universe for which the space time metric is rather simple:

$$ds^2 = dt^2 - a^2(t) d\vec{x}^2 \quad (62)$$

the time and spatial variables  $t, \vec{x}$  in the above metric are called comoving time and spatial distance respectively and have the interpretation of being the time and distance measured by an observer locally at rest with respect to the Hubble flow. At this point we must note that *physical distances* are given by  $\vec{l}_{phys}(t) = a(t) \vec{x}$ . An important concept is that of causal (particle) horizons: events that cannot be connected by a light signal are causally disconnected. Since light travels on null geodesics  $ds^2 = 0$  the maximum *physical* distance that can be reached by a light signal at time  $t$  is given by

$$d_H(t) = a(t) \int_0^t \frac{dt'}{a(t')} \quad (63)$$

It will prove convenient to change coordinates to *conformal time* by defining a conformal time variable

$$\eta = \int_0^t \frac{dt'}{a(t')} \Rightarrow ds^2 = C^2(\eta) (d\eta^2 - d\vec{x}^2) ; \quad C(\eta) = a(t(\eta)) \quad (64)$$

in terms of which the causal horizon is simply given by  $d_H(\eta) = C(\eta) \eta$  and physical distances as  $\vec{x}_{phys} = C(\eta) \vec{x}$ . This metric is of the same form as that of Minkowski space time. For energies well below the Planck scale  $M_{Pl} \approx 10^{19}$  Gev gravitation is well described by *classical* General Relativity and the Einstein equations:

$$R^{\mu\nu} - \frac{1}{2} g^{\mu\nu} R = \frac{8\pi}{3M_{Pl}^2} T^{\mu\nu} \quad (65)$$

where we have been cavalier and set  $c = 1$  (as well as  $\hbar = 1$ ).  $R^{\mu\nu}$  is the Ricci tensor,  $R$  the Ricci scalar and  $T^{\mu\nu}$  the matter field energy momentum tensor. The above equation is classical but one seeks to understand the dynamics of the Early Universe in terms of a *quantum field theory* that describes particle physics, thus the question: what is exactly the energy momentum tensor?, in Einstein's equations it is a classical object, but in QFT it is an operator. The answer to this question is: gravity is classical, fields are quantum mechanical, but  $T^{\mu\nu} \rightarrow \langle T^{\mu\nu} \rangle$ , i.e. it is the expectation value of a *quantum mechanical operator in a quantum mechanical state*. This quantum mechanical state, either pure or mixed is described by a wave-function(al) or a density matrix whose time evolution is dictated by the quantum equations of motion: the Schrödinger equation for the wave functions or the quantum Liouville equation for a density matrix. Consistency with the postulate of homogeneity and isotropy requires that the expectation value of the energy momentum tensor must have the fluid form and in the rest frame of the fluid takes the form  $\langle T^{\mu\nu} \rangle = \text{diagonal}(\rho, p, p, p)$  with  $\rho$  the energy density and  $p$  the pressure. The time and spatial components of Einstein's equations lead to the Friedman equation

$$\frac{\dot{a}^2(t)}{a^2(t)} = \frac{8\pi}{3M_{Pl}^2} \rho(t) \quad (66)$$

$$2\frac{\ddot{a}(t)}{a(t)} + \frac{\dot{a}^2(t)}{a^2(t)} = -\frac{8\pi}{M_{Pl}^2} p(t) \quad (67)$$

Combining these two equations one arrives at a simple and intuitive equation which is reminiscent of the first law of thermodynamics:

$$\frac{d}{dt}(\rho a^3(t)) = -p \frac{da^3(t)}{dt} \Rightarrow \dot{\rho} + 3\frac{\dot{a}}{a}(\rho + p) = 0 \quad (68)$$

The alternative form shown on the right hand side of (68) is the *covariant conservation of energy*. Since the physical volume of space is  $V_0 a^3(t)$  (with  $V_0$  the comoving volume) the above equation is recognized as  $dU = -p dV$  which is the first law of thermodynamics for *adiabatic* processes. To close the set of equations and obtain the dynamics we need an equation of state  $p = p(\rho)$ : two very relevant cases are: i) radiation dominated (RD) with  $p = \rho/3$  and matter dominated (MD)  $p = 0$  (dust) Universes. In our study we will focus on the RD case. The equation of state for RD is that for blackbody radiation for which the entropy is  $S = CVT^3$  (with  $C$  a constant). Since  $V(t) = V_0 a^3(t)$  is the physical volume, the equation (68) which dictates adiabatic (isoentropic) expansion leads to a time dependence of the temperature:  $T(t) = T_0/a(t)$ . Now the cooling is done by the expansion of the Universe and a phase transition will occur when the Universe cools below the critical temperature for a given theory. For the GUT transition  $T_c \approx 10^{16} \text{ GeV} \approx 10^{29} \text{ K}$ , for the EW transition  $T_c \approx 100 \text{ GeV} \approx 10^{15} \text{ K}$ . Returning now back to the large  $N$  study of the dynamics of phase transitions, we can include the effect of cooling by the expansion of the Universe by replacing the time dependent mass term  $m^2(t)$  in (43) by

$$m^2(t) = m_0^2 \left[ \frac{T^2(t)}{T_c^2} - 1 \right] \quad ; \quad T(t) = \frac{T_i}{a(t)} \quad (69)$$

This form is consistent with the Landau-Ginzburg description including the time dependence of the temperature via the isentropic expansion of the Universe, but perhaps more importantly it can be proven in a detailed manner from the self-consistent renormalization of the mass in an expanding Universe[34]. Thus the large  $N$  limit in a RD FRW cosmology will be studied by using the potential (43) but with the time dependent mass given by (69).

## 4.2 Large $N$ in Radiation (RD) and Matter (MD) dominated FRW Cosmology

The large  $N$  limit is again implemented via the Hartree-like factorization (44) performing the spatial Fourier transforms of the fields and their canonical momenta and including the proper scale factors, the Hamiltonian now becomes[34]



$$H(t) = \sum_k \left\{ \frac{1}{2a^3(t)} \vec{\Pi}_{\vec{k}} \cdot \vec{\Pi}_{-\vec{k}} + W_k^2(t) \vec{\Phi}_{\vec{k}} \cdot \vec{\Phi}_{-\vec{k}} \right\} \quad (70)$$

$$W_k^2(t) = \frac{k^2}{a^2} + m^2(t) + \frac{\lambda}{2N} \langle \vec{\Phi}_{\vec{k}} \cdot \vec{\Phi}_{-\vec{k}} \rangle \quad (71)$$

where now the expectation value is in terms of a *density matrix*  $\rho[\Phi(\vec{\tau}), \vec{\Phi}(\vec{\tau}); t]$  since we are considering the case of a thermal ensemble as the initial state.

We propose the following Gaussian ansatz for the functional density matrix elements in the Schrödinger representation[34]

$$\rho[\Phi, \vec{\Phi}, t] = \prod_{\vec{k}} \mathcal{N}_k(t) \exp \left\{ -\frac{A_k(t)}{2} \vec{\Phi}_{\vec{k}} \cdot \vec{\Phi}_{-\vec{k}} + \frac{A_k^*(t)}{2} \vec{\tilde{\Phi}}_{\vec{k}} \cdot \vec{\tilde{\Phi}}_{-\vec{k}} + B_k(t) \vec{\Phi}_{\vec{k}} \cdot \vec{\tilde{\Phi}}_{-\vec{k}} \right\} \quad (72)$$

This form of the density matrix is dictated by the hermiticity condition  $\rho^\dagger[\Phi, \vec{\Phi}, t] = \rho^*[\vec{\Phi}, \Phi, t]$ ; as a result of this,  $B_k(t)$  is real. The kernel  $B_k(t)$  determines the amount of mixing in the density matrix, since if  $B_k = 0$ , the density matrix corresponds to a pure state because it is a wave functional times its complex conjugate. The kernels  $A_k(0)$  ;  $B_k(0)$  are chosen such that the initial density matrix is thermal with a temperature  $T_i > T_c$ [34]. Following the same steps as in Minkowski space time, the time evolution of this density matrix can be found in terms of a set of mode functions  $\phi_k(t)$  that obey the following equations of motion and self-consistency condition

$$\ddot{\phi}_k(t) + 3 \frac{\dot{a}}{a} \dot{\phi}_k(t) + \left[ \frac{k^2}{a^2(t)} + m^2(t) \right] \phi_k(t) = 0 \quad (73)$$

$$m^2(t) = m_0^2 \left[ \frac{T_i^2}{T_c^2 a^2(t)} - 1 \right] + \frac{\lambda}{2} \int \frac{d^3 k}{(2\pi)^3} |\phi_k(t)|^2 \coth \frac{W_k(0)}{2T_i} . \quad (74)$$

This equations can be cast in a more familiar form by changing coordinates to conformal time (see eq. (64)) and (conformally) rescaling the mode functions  $\phi_k(t) = f_k(\eta)/C(\eta)$  to obtain the following equations for the conformal time mode functions  $f_k(\eta)$  in an RD FRW cosmology

$$f_k''(\eta) + [k^2 + C^2(\eta)M^2(\eta)] f_k(\eta) = 0 \quad (75)$$

$$M^2(\eta) = m_0^2 \left[ \frac{T_i^2}{T_c^2 C^2(\eta)} - 1 \right] + \frac{\lambda}{2} \int \frac{d^3 k}{(2\pi)^3} \left[ \frac{|f_k(\eta)|^2}{C^2(\eta)} \coth \frac{W_k(0)}{2T_i} \right] - \frac{C''(\eta)}{C^3(\eta)} \quad (76)$$

where primes now refer to derivatives with respect to conformal time. For RD and MD FRW

$$C(\eta) = 1 + \frac{\eta}{2} ; \quad C'''(\eta) = 0 \quad \text{for RD} \quad (77)$$

$$C(\eta) = (1 + \frac{\eta}{4})^2 ; \quad C'''(\eta) = 1/8 \quad \text{for MD} \quad (78)$$

$$(79)$$

(in units of  $m_0^{-1}$  which is the only dimensionful variable). The above equations of motion now have a form analogous to those in the case of Minkowski space-time,

As the temperature falls below the critical the effective squared mass term becomes negative and spinodal instabilities trigger the process of phase ordering. This results in that the quantum fluctuations quantified by  $\langle \vec{\Phi}^2 \rangle$  grow exponentially. These spinodal instabilities make the self-consistent field grow at early times and tends to overcome the negative sign of the squared mass, eventually reaching an asymptotic regime in which the total effective mass  $M^2(\eta)$  vanishes.

Again this behavior determines that the fluctuations are sampling the equilibrium broken symmetry minima of the initial potential, i.e.  $\langle \vec{\Phi}^2 \rangle \rightarrow \frac{2Nm_0^2}{\lambda}$ .

Although, just as in Minkowski space-time the effective mass vanishes asymptotically, the non-equilibrium evolution is rather *different*. We find numerically[38] that asymptotically the effective

mass term behaves as

$$C^2(\eta)M^2(\eta) \xrightarrow{\eta \rightarrow \infty} -15/4\eta^2 \text{ for RD} \quad (80)$$

$$C^2(\eta)M^2(\eta) \xrightarrow{\eta \rightarrow \infty} -35/4\eta^2 \text{ for MD} \quad (81)$$

Fig. 6 displays  $C^2(\eta)M^2(\eta)$  as a function of conformal time for the case of  $T_i/T_c = 1.1$  with  $T_c \propto m_0/\sqrt{\lambda}$ [34, 44] for RD.

We see that at very early time the mass is positive, reflecting the fact that the initial state is in equilibrium at an initial temperature larger than the critical. As time evolves the temperature is red-shifted and cools and at some point the phase transition occurs, when the mass vanishes and becomes negative.

Figure 7 displays  $\frac{\lambda}{2Nm_0^2} \langle \tilde{\Phi}^2 \rangle(\eta)$  vs.  $\eta$  in units of  $m_0^{-1}$  for  $\frac{T_i}{T_c} = 3$ ,  $g = 10^{-5}$  for an R.D. Universe. Clearly at large times the non-equilibrium fluctuations probe the broken symmetry states.

This particular asymptotic behavior of the mass determines that the mode functions  $f_k(\eta)$  grow as  $\eta^{5/2}$  for RD and for  $\eta^{7/2}$   $k < 1/\eta$  and oscillate in the form  $e^{\pm i k \eta}$  for  $k > 1/\eta$ . This behavior is confirmed numerically[38]. We find both analytically and numerically that asymptotically the mode functions are of the *scaling form*

$$f_k(\eta) = A \eta^{\frac{5}{2}} \frac{J_2(k\eta)}{(k\eta)^2} \text{ for RD} \quad (82)$$

$$f_k(\eta) = B \eta^{\frac{7}{2}} \frac{J_3(k\eta)}{(k\eta)^3} \text{ for MD} \quad (83)$$

Where  $A$  and  $B$  are numerical constants and  $J_{2,3}(x)$  are Bessel functions.

Figure 8 displays  $\eta^{-5}|f_k(\eta)|^2$  as a function of the scaling variable  $k\eta$  revealing the scaling behavior in RD, a similar behavior emerges for MD[38].

It is remarkable that this is exactly the same scaling solution found in the *classical* non-linear sigma model in the large  $N$  limit and that describes the collapse of textures[32], and also within the context of TDGL equations in the large  $N$  limit applied to cosmology[33].

The growth of the long-wavelength modes and the oscillatory behavior of the short wavelength modes again results in that the peak of the structure factor  $S(k, \eta) = |f_k(\eta)|^2 \propto C^2(\eta)\eta^3 g(k\eta)$  moves towards longer wavelengths and the maximum amplitude increases. This is the equivalent of coarsening and the onset of a condensate.

Although quantitatively different from Minkowsky space time, the qualitative features are similar. Asymptotically the non-equilibrium dynamics results in the formation of a non-perturbative condensate of long-wavelength Goldstone bosons. We can now compute the pair correlation function  $C(r, \eta)$  from the mode functions solutions to (75) and find that it is cutoff by causality at  $r = 2\eta$ . The correlation function computed with the mode functions in the asymptotic regime agrees perfectly with that computed from the asymptotic form given by (82). The correlation function is depicted in Fig. 9 for two different (conformal) times.

The scaling form of the pair correlation function is

$$C(r, \eta) \propto \eta^2 \chi_{RD}(r/2\eta) \text{ for RD} \quad C(r, \eta) \propto \eta^4 \chi_{MD}(r/2\eta) \text{ for MD} ,$$

where  $\chi_{RD}(x)$  and  $\chi_{MD}(x)$  are hump-shaped functions as shown in fig. 9.

Clearly a *dynamical* length scale  $L(\eta) = \eta$  emerges as a consequence of causality, much in the same manner as in Minkowsky space time. The *physical* dynamical correlation length is therefore given by  $\xi_{phys}(\eta) = C(\eta)L(\eta) = d_H(t)$ , that is the correlated domains grow again at the speed of light and their size is given by the causal horizon. The interpretation of this phenomenon is that within one causal horizon there is one correlated domain, inside which the mean square root fluctuation of the field is approximately the value of the equilibrium minima of the tree level potential, this is clearly consistent with Kibble's original observation[2]. Inside this domain there is a non-perturbative condensate of Goldstone bosons[38].

Thus we have seen that the phenomenon of scaling, coarsening and the onset of condensates during the non-equilibrium dynamics of phase ordering is a *universal* feature of the process of

phase ordering. The non-perturbative large  $N$  limit has allowed a clear comparison between the phenomenological description in condensed matter based on the TDGL, and the microscopic quantum field theoretical description in Minkowski and FRW space-times.

## 5 Conclusions and looking ahead

In this lectures we have discussed the multidisciplinary nature of the problem of phase ordering kinetics and non-equilibrium aspects of symmetry breaking. Main ideas from condensed matter were discussed and presented in a simple but hopefully illuminating framework and applied to the rather different realm of phase transitions in quantum field theory as needed to understand cosmology and particle physics. In particular we have emphasized *robust* features of the process of phase ordering kinetics: early stages dominated by spinodal instabilities and the growth of correlated regions, the emergence of a dynamical correlation length that determines the size of the correlated regions as a function of time and *dynamical scaling* at long times. The phenomenon of coarsening is a result of this scaling behavior and is reflected in that the peak in the power spectrum moves towards longer wavelengths, and asymptotically long times results in a 'Bragg peak' that signals the onset of macroscopic ordered phases and condensates. The study of condensed matter systems was in terms of the phenomenologically succesful Time Dependent Landau Ginzburg theory which is purely dissipative and for which there is no first principles derivation from a microscopic theory in general.

We then passed onto the study of the dynamical evolution out of equilibrium in quantum field theories both in Minkowsky and FRW space-times by providing a consistent *non-perturbative* framework to study the time evolution of an initially prepared density matrix.

The large  $N$  approximation has provided a bridge that allows to cross from one field to **another** and borrow many of the ideas that had been tested both theoretically and experimentally in condensed matter physics. There are, however, major differences between the condensed matter and particle physics-cosmology applications that require a very careful treatment of the quantum field theory that cannot be replaced by simple arguments. The large  $N$  approximation in field theory provides a robust, consistent non-perturbative framework that allows the study of phase ordering kinetics and dynamics of symmetry breaking in a controlled and consistently implementable framework, it is renormalizable, respects all symmetries and can be improved in a well defined manner. This scheme extracts cleanly the non-perturbative behavior, the quantum to classical transition and allows to quantify in a well defined manner the emergence of classical stochastic behavior arising from non-perturbative physics. The emergence of scaling and a dynamical correlation length, coarsening and the onset of non-perturbative condensates are robust features of the dynamics and the Kibble-Zurek scenario describes fairly well the general features of the dynamics, albeit the details require careful study, both analytically and numerically.

We have emphasized that this study has very definite potential experimental implications, in QCD if the chiral phase transition occurs out of equilibrium in ultrarelativistic heavy ion collisions leads to the possibility of formation of disoriented chiral condensates that are described in the same manner as ordering domains in condensed matter. These condensates have a very distinct hallmark in that they lead to a very different ratio of neutral to charged pions, this property can be measured on an event by event basis with the detectors at the forthcoming heavy ion colliders.

In cosmology the process of formation of ordered regions that grow after a rapid phase transition, the emergence of scaling and a dynamical length scale and coarsening of these domains lead to a definite prediction of a 'red' power spectrum on scales that have re-entered the causal horizon right after recombination. These are the scales that contribute to the temperature anisotropies measured by COBE and the forthcoming cosmological experiments. Therefore the study of the dynamics of symmetry breaking out of equilibrium in quantum field theory directly bears on experimental possibilities in a wide range of energies both in accelerator and cosmological experiments and is therefore an endeavour that must be pursued vigorously.

Of course this is just the beginning, we expect a wealth of important phenomena to be revealed beyond the large  $N$ , such as the approach to equilibrium, the emergence of other time scales

associated with a hydrodynamic description of the evolution at late times and a more careful understanding of the reheating process and its influence on cosmological observables. Although within very few years the wealth of observational data will provide a more clear picture of the cosmological fluctuations, it is clear that the program that pursues a fundamental understanding of the underlying physical mechanisms will continue seeking to provide a consistent microscopic description of the dynamics of particle physics and cosmological phase transitions.

## 6 Acknowledgements:

D. B. thanks T. Kibble, W. Zurek and R. Durrer for illuminating conversations, the N.S.F for partial support through grant awards: PHY-9605186 and INT-9815064 and LPTHE (University of Paris VI and VII) for warm hospitality, H. J. de Vega thanks the Dept. of Physics at the Univ. of Pittsburgh for hospitality. We thank NATO for partial support.

## References

- [1] For a comprehensive review of the status of theory and experiment see: Proceedings of the 'D. Chalonge' School in Astrofundamental Physics at Erice, edited by N. Sánchez and A. Zichichi, 1996 World Scientific publisher and 1997, Kluwer Academic publishers. In particular the contributions by G. Smoot, A. N. Lasenby and A. Szalay.,
- [2] T. W. B. Kibble, J. Phys. A 9, 1387 (1976). M. B. Hindmarsh and T.W.B. Kibble, Rep. Prog. Phys. 58:477 (1995).  
A. Vilenkin and E.P.S. Shellard, 'Cosmic Strings and other Topological Defects', Cambridge Monographs on Math. Phys. (Cambridge Univ. Press, 1994).
- [3] For thorough reviews of standard and inflationary cosmology see: E. W. Kolb and M. S. Turner, *The Early Universe* (Addison Wesley, Redwood City, C.A. 1990). A. Linde, *Particle Physics and Inflationary Cosmology*, (Harwood Academic Pub. Switzerland, 1990). R. Brandenberger, Rev. of Mod. Phys. 57,1 (1985); Int. J. Mod. Phys. A2, 77 (1987).
- [4] For more recent reviews see: M. S. Turner, astro-ph-9703197;astro-ph-9703196;astro-ph-9703174;astro-ph-9703161; astro-ph-9704062.
- [5] See for example: K. Rajagopal in 'Quark Gluon Plasma 2', (Ed. R. C. Hwa, World Scientific, 1995).
- [6] For a thorough discussion of phase transitions see: N. Goldenfeld, 'Lectures on Phase Transitions and the Renormalization Group', (Addison-Wesley, 1992).
- [7] A. J. Bray, Adv. Phys. 43, 357 (1994).
- [8] J. S. Langer in 'Solids far from Equilibrium', Ed. C. Godrèche, (Cambridge Univ. Press 1992); J. S. Langer in 'Far from Equilibrium Phase Transitions', Ed. L. Garrido, (Springer-Verlag, 1988); J. S. Langer in 'Fluctuations, Instabilities and Phase Transitions', Ed. T. Riste, Nato Advanced Study Institute, Geilo Norway, 1975 (Plenum, 1975).
- [9] G. Mazenko in in 'Far from Equilibrium Phase Transitions', Ed. L. Garrido, (Springer-Verlag, 1988).
- [10] C. Castellano and M. Zannetti, cond-mat/9807242; C. Castellano, F. Corberi and M. Zannetti, Phys. Rev. E56, 4973 (1997); F. Corberi, A. Coniglio and M. Zannetti, Phys. Rev. E51, 5469 (1995).
- [11] W. H. Zurek, Nature 317, 505 (1985); Acta Physica Polonica B24, 1301 1993); Phys. Rep. 276, 4 (1996).

- [12] W. I. Goldburg and J. S. Huang, in 'Fluctuations, Instabilities and Phase Transitions', Ed. T. Riste, Nato Advanced Study Institute, Geilo Norway, 1975 (Plenum, 1975); J. S. Huang, W. I. Goldburg and M. R. Moldover, *Phys. Rev. Lett.* **34**, 639 (1975).
- [13] For a nicely written recent review on the dynamics of phase transition see: A. Gill, 'Contemporary Physics', vol. 39, number 1, pages 13-47 (1998).
- [14] I. Chuang, R. Durrer, N. Turok and B. Yurke, *Phys. Rev. Lett.* **66**, 2472 (1990).
- [15] M. Bowick, L. Chandar, E. Schiff and A. Srivastava, *Science* **263**, 943 (1994).
- [16] P. C. Hendry, N. S. Lawson, R. A. M. Lee, P. V. E. McClintock and C. D. H. Williams, *Nature*, **368**, 315 (1994).
- [17] V.M.H. Ruutu et. al., cond-mat/9512117. Y. M. Bunkov and O. D. Timofeevskaya, cond-mat/9706004.
- [18] For recent reviews on the QCD phase transitions and aspects of relativistic heavy ion collisions see for example: J. W. Harris and B. Muller, *Annu. Rev. Nucl. Part. Sci.* **46**, 71 (1996). B. Muller in *Particle Production in Highly Excited Matter*, Eds. H.H. Gutbrod and J. Rafelski, NATO ASI series B, vol. 303 (1993). B. Muller, *The Physics of the Quark Gluon Plasma* Lecture Notes in Physics, Vol. 225 (Springer-Verlag, Berlin, Heidelberg, 1985); K. Rajagopal in 'Quark-Gluon Plasma 2', Ed. by R. C. Hwa (World Scientific, Singapore) (1995); H. Meyer-Ortmanns, *Rev. of Mod. Phys.* **68**, 473 (1996). C-Y Wong, 'Introduction to High-Energy Heavy Ion Collisions', (World Scientific, 1994).
- [19] J. D. Bjorken, *Phys. Rev. D* **27**, 140 (1982).
- [20] K. Rajagopal and F. Wilczek, *Nucl. Phys. B* **399**, 395 (1993); K. Rajagopal and F. Wilczek, *Nucl. Phys. B* **404**, 577 (1993).
- [21] A. A. Anselm and M. G. Ryskin, *Phys. Lett. B* **266**, (1991) 482; J. D. Bjorken, K. L. Kowalski and C. C. Taylor, SLAC Report No. SLAC-PUB-6109 (unpublished); J. - P. Blaizot and A. Krzywicki, *Phys. Rev. D* **46**, 1992 (246); J. D. Bjorken, *Int. J. Mod. Phys. A* **7**, (1992) 4189; J. D. Bjorken, *Acta Physica Polonica B* **23**, (1992) 561; K. L. Kowalski and C. C. Taylor, 'Disoriented Chiral Condensate: A White Paper for the Full Acceptance Detector' CWRU report 92- he-ph/9211282 (unpublished); J. D. Bjorken, K.L. Kowalski and C. C. Taylor, 'Baked Alaska', Proceedings of Les Rencontres de Physique del Valle d'Aoste, La Thuile (1993); (SLAC PUB 6109). G. Amelino-Camelia, J. D. Bjorken, S. E. Larsson, *Phys. Rev. D* **56** (1997) 6942; J. D. Bjorken, *Acta Phys. Polon. B* **28** (1997) 2773; A. Anselm, *Phys. Lett. B* **217**, 169 (1989).
- [22] L. T. Baradzei et. al. *Nucl. Phys. B* **370**, (1992) 365.
- [23] S. Gavin, A. Gocksch and R. D. Pisarski, *Phys. Rev. Lett.* **72**, 2143 (1994); S. Gavin and B. Muller, *Phys. Lett. B* **329**, 486 (1994); Z. Huang and X.-N. Wang, *Phys. Rev. D* **49**, 4335 (1994); Z. Huang, M. Suzuki and X.-N. Wang, *Phys. Rev. D* **50**, 2277 (1994); Z. Huang and M. Suzuki, *Phys. Rev. D* **53**, 891 (1996); M. Asakawa, Z. Huang and X. N. Wang, *Phys. Rev. Lett.* **74**, 3126 (1995); J. Randrup, *Nucl. Phys. A* **616** (1997) 531; J. Randrup, *Phys. Rev. Lett.* **77** (1996) 1226.
- [24] D. Boyanovsky, H. J. de Vega and R. Holman, *Phys. Rev. D* **51**, (1995) 734; F. Cooper, Y. Kluger, E. Mottola and J. P. Paz, *Phys. Rev. D* **51**, (1995) 2377. Y. Kluger, F. Cooper, E. Mottola, J. P. Paz and A. Kovner, *Nucl. Phys. A* **590**, (1995) 581.
- [25] WA98 Collaboration, (M. M. Aggarwal et. al.) *Phys. Lett. B* **420**, (1998) 169.
- [26] J. Streets, hep-ex/9608012; T. C. Brooks et. al. *Phys. Rev. D* **55**, (1997), 5667; M. E. Convery, hep-ex/9801020.

- [27] see the RHIC project page with detailed description of the physics capabilities of STAR and PHENIX at <http://www.rhic.bnl.gov>
- [28] see the Castor project page at the Alice web page, <http://www1.cern.ch/ALICE/projects.html>.
- [29] D. Boyanovsky, D. Cormier, H. J. de Vega, R. Holman and S. P. Kumar, Phys. Rev. D. 57, 2166 (1998); D. Boyanovsky, D. Cormier, H. J. de Vega and R. Holman, Phys. Rev. D 55 (1997) 3373.
- [30] D. Boyanovsky, H. J. de Vega and R. Holman, hep-ph/9903534 and in preparation.
- [31] C. L. Bennet, M. S. Turner and M. White, Physics Today, 50 NOV 32, (1997) and references therein.
- [32] N. Turok and D. N. Spergel, Phys. Rev. Lett. 66, 3093 (1991); D. N. Spergel, N. Turok, W. H. Press and B. S. Ryden, Phys. Rev. D 43, 1038 (1991).
- [33] J. A. N. Filipe and A. J. Bray, Phys. Rev. E 50, 2523 (1994); J. A. N. Filipe, (Ph. D. Thesis, 1994, unpublished).
- [34] D. Boyanovsky, H. J. de Vega and R. Holman, Phys. Rev. D 49, 2769 (1994); D. Boyanovsky, D. Cormier, H. J. de Vega, R. Holman et S. Prem Kumar, Phys. Rev. D 57, 2166, (1998), (and references therein).
- [35] D. Boyanovsky, H. J. de Vega, R. Holman, D.-S. Lee and A. Singh, Phys. Rev. D 51, 4419 (1995). D. Boyanovsky, H. J. de Vega and R. Holman, Proceedings of the Second Paris Cosmology Colloquium, Observatoire de Paris, June 1994, pp. 127-215, H. J. de Vega and N. Sánchez, Editors (World Scientific, 1995); Advances in Astrofundamental Physics, Erice Chalonge School, N. Sánchez and A. Zichichi Editors, (World Scientific, 1995). D. Boyanovsky, H. J. de Vega, R. Holman and J. Salgado, Phys. Rev. D 54, 7570 (1996); D. Boyanovsky, D. Cormier, H. J. de Vega, R. Holman, A. Singh, M. Srednicki; Phys. Rev. D 56 (1997) 1939. D. Boyanovsky, H. J. de Vega and R. Holman, Vth. Erice Chalonge School, Current Topics in Astrofundamental Physics, N. Sánchez and A. Zichichi Editors, World Scientific, 1996, p. 183-270. D. Boyanovsky, M. D'Attanasio, H. J. de Vega, R. Holman and D. S. Lee, Phys. Rev. D 52, 6805 (1995). D. Boyanovsky, H. J. de Vega, R. Holman and J. Salgado, Phys. Rev. D 57, 7388 (1998).
- [36] D. Boyanovsky, H. J. de Vega, R. Holman and J. Salgado, hep-ph/9811273, to appear in Phys. Rev. D.
- [37] F. Cooper, S. Habib, Y. Kluger, E. Mottola, Phys. Rev. D 55 (1997), 6471. F. Cooper, S. Habib, Y. Kluger, E. Mottola, J. P. Paz, P. R. Anderson, Phys. Rev. D 50, 2848 (1994). F. Cooper, Y. Kluger, E. Mottola, J. P. Paz, Phys. Rev. D 51, 2377 (1995); F. Cooper and E. Mottola, Mod. Phys. Lett. A 2, 635 (1987); F. Cooper and E. Mottola, Phys. Rev. D 36, 3114 (1987); F. Cooper, S.-Y. Pi and P. N. Stancioff, Phys. Rev. D 34, 3831 (1986).
- [38] D. Boyanovsky and H. J. de Vega, in preparation.
- [39] L. F. Cugliandolo and D. S. Dean, J. Phys. A 28, 4213 (1995); *ibid* L453, (1995); L. F. Cugliandolo, J. Kurchan and G. Parisi, J. Physique (France) 4, 1641 (1994).
- [40] See D. Boyanovsky, R. Holman and H. J. de Vega in [24], and the first reference in [37].
- [41] Relaxing the assumption of an instantaneous quench and allowing for a time dependence of the cooling mechanism has been recently studied by M. Bowick and A. Momen, hep-ph/9803284.
- [42] E. J. Weinberg and A. Wu, Phys. Rev. D 36, 2474 (1987); A. Guth and S.-Y. Pi, Phys. Rev. D 32, 1899 (1985).

- [43] D. Boyanovsky and H. J. de Vega, *Phys. Rev. D***47**, 2343 (1993); D. Boyanovsky *Phys. Rev. E***48**, 767 (1993).
- [44] D. Boyanovsky, D.-S. Lee and A. Singh, *Phys. Rev. D***48**, 800 (1993).
- [45] G. Karra and R.J.Rivers, *Phys.Lett. B***414** (1997), 28; R.J.Rivers, 3rd. Colloque Cosmologie, Observatoire de Paris, June 1995, p. 341 in the Proceedings edited by H J de Vega and N. Sánchez, World Scientific. A.J. Gill and R.J. Rivers, *Phys.Rev. D***51** (1995), 6949; G.J. Cheetham, E.J. Copeland, T.S. Evans, R.J. Rivers, *Phys.Rev.D***47** (1993),5316.
- [46] 'Defect Formation and Critical Dynamics in the Early Universe', G. J. Stephens, E. A. Calzetta, B. L. Hu, S. A. Ramsey, gr-qc/9808059 (1998). 'Counting Defects in an Instantaneous Quench', D. Ibaceta and E. Calzetta, hep-ph/9810301 (1998).

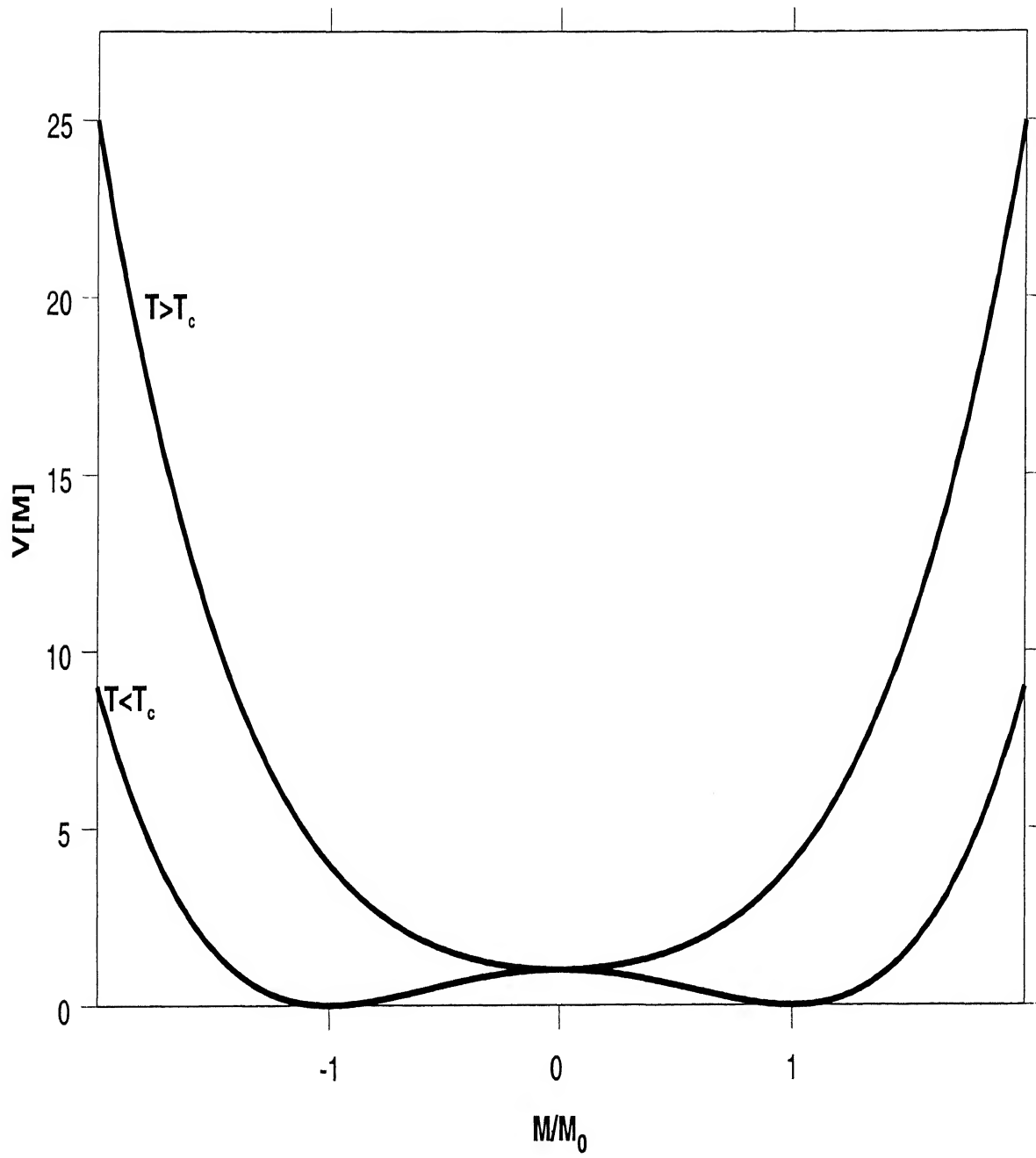


Figure 1:  $V[M]$  vs.  $M$ , for  $T > T_c$  and  $T < T_c$



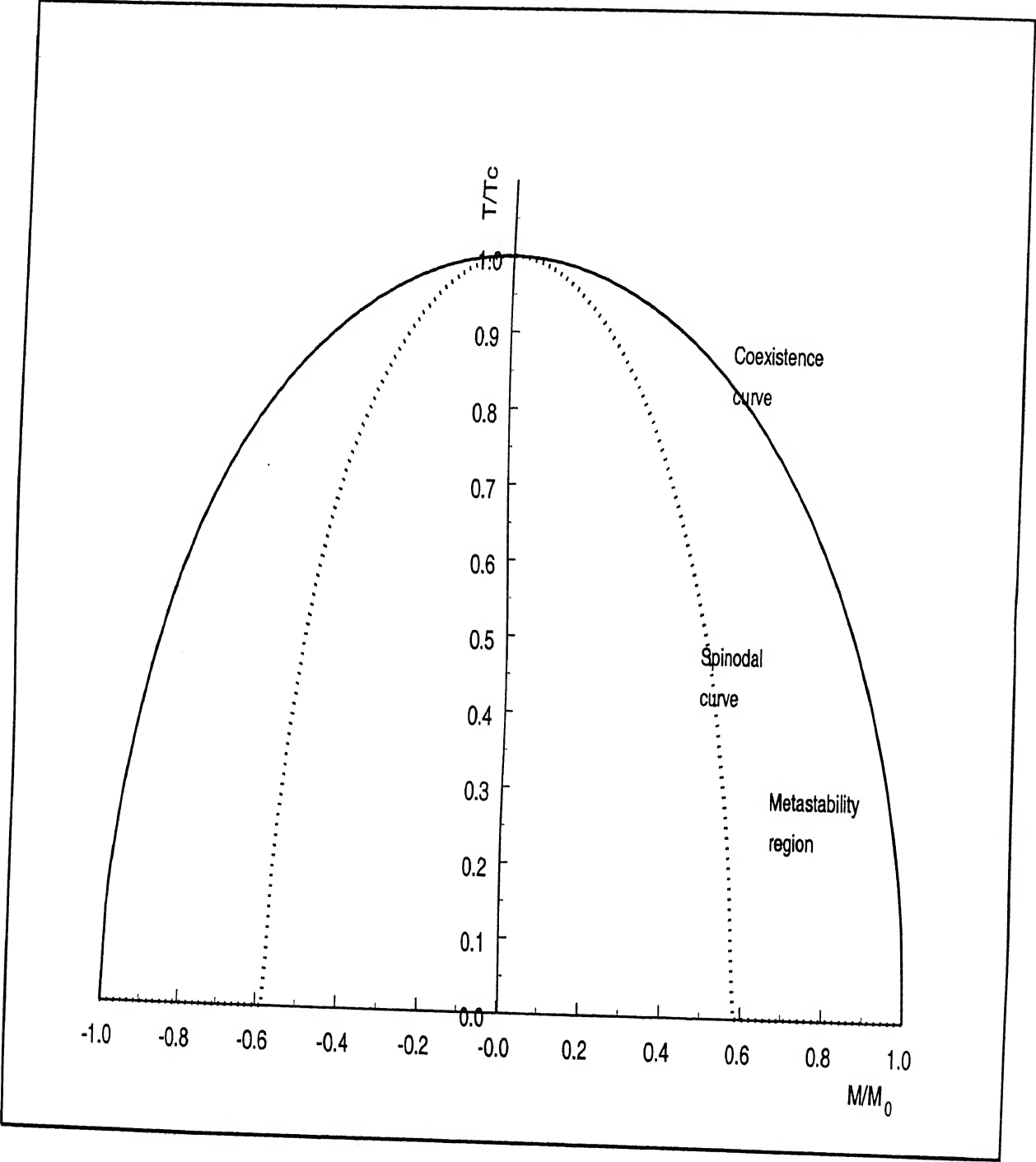


Figure 2: Classical spinodal and coexistence curves for the potential  $V[M]$  in (1)

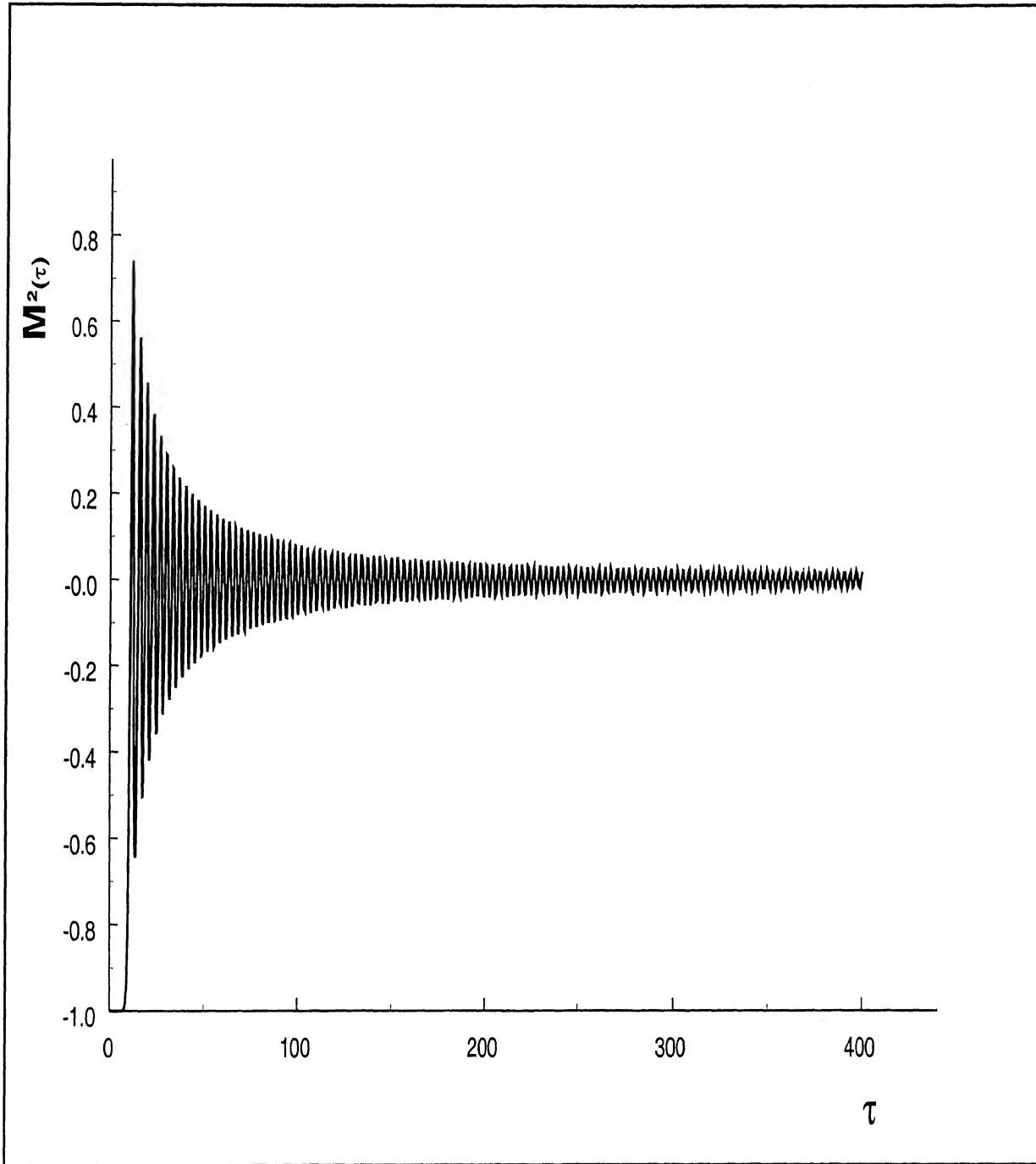


Figure 3:  $\mathcal{M}^2(\tau)$  vs.  $\tau$ ,  $g = 10^{-7}$

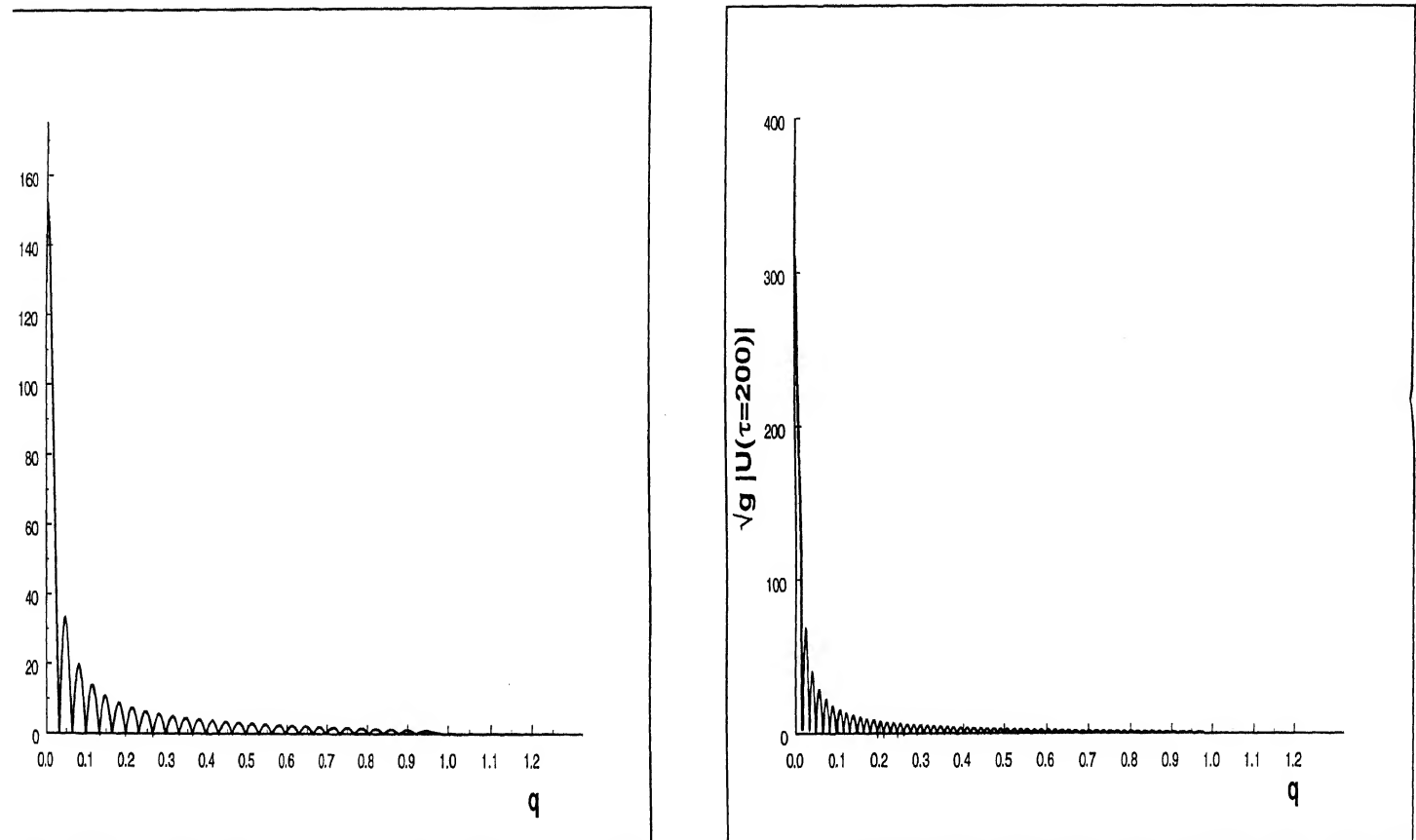


Figure 4:  $g|\phi_k(\tau = 100, 200)|^2$  vs.  $q = k/|m_0|$ ,  $g = 10^{-7}$

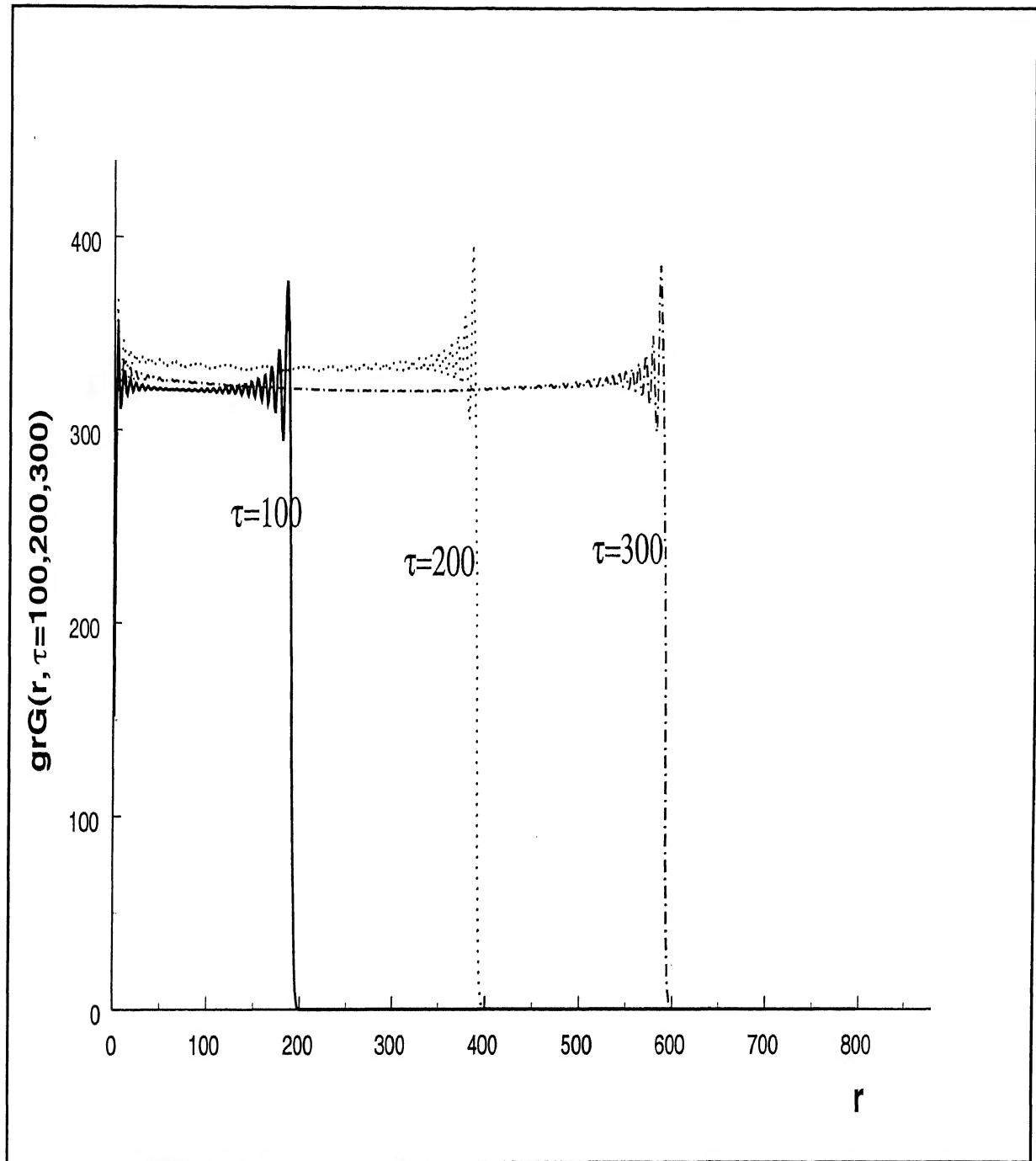


Figure 5:  $gr C(r, \tau)$  vs  $r/|m_0|$  for  $t/|m_0| = 100, 200, 300$  for  $g = 10^{-7}$ .

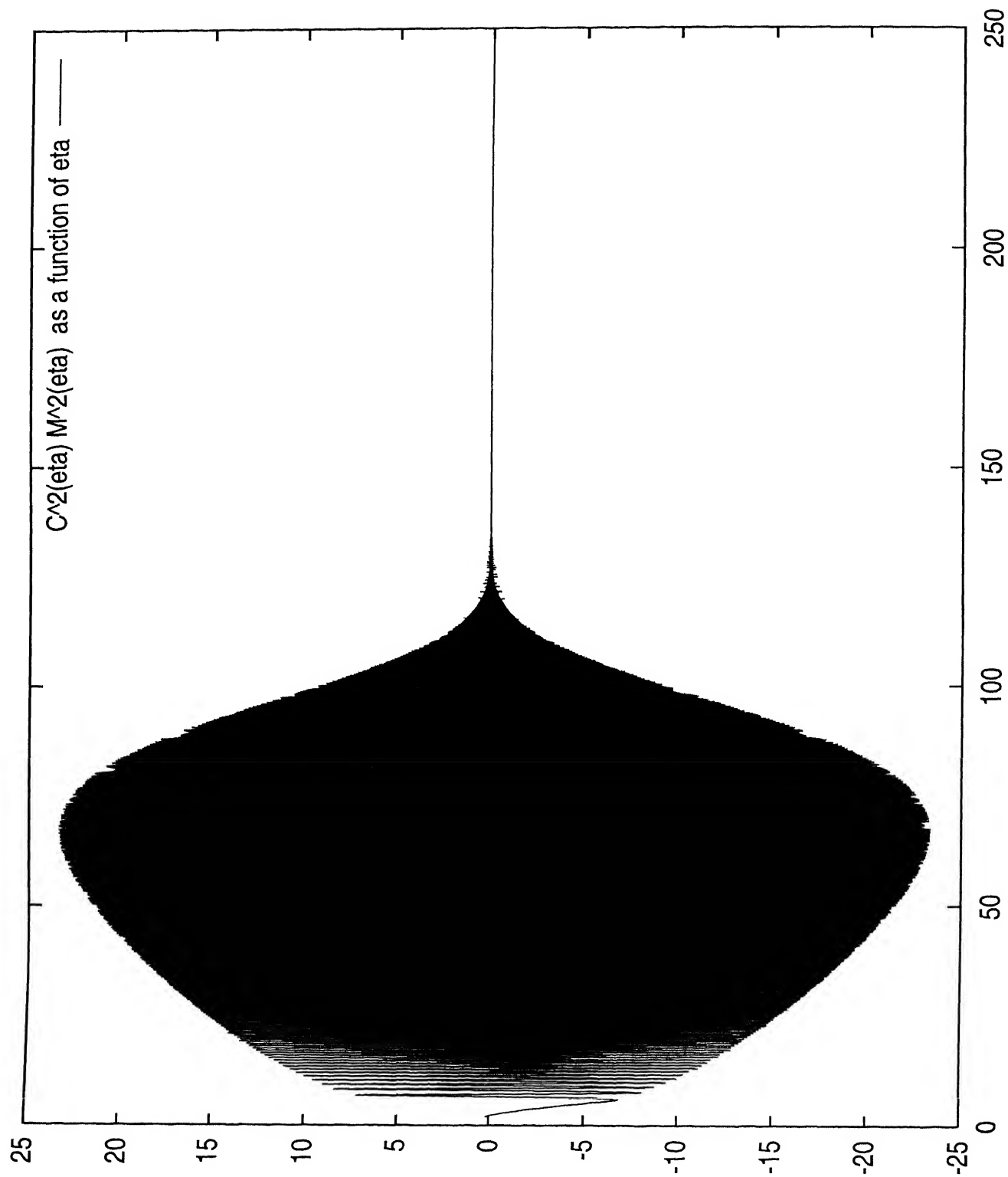


Figure 6:  $C^2(\eta)M^2(\eta)$  vs.  $\eta$ (conformal time in units of  $m_0^{-1}$ ) for  $\frac{T_i}{T_c} = 3, g = 10^{-5}$ . R.D. Universe.

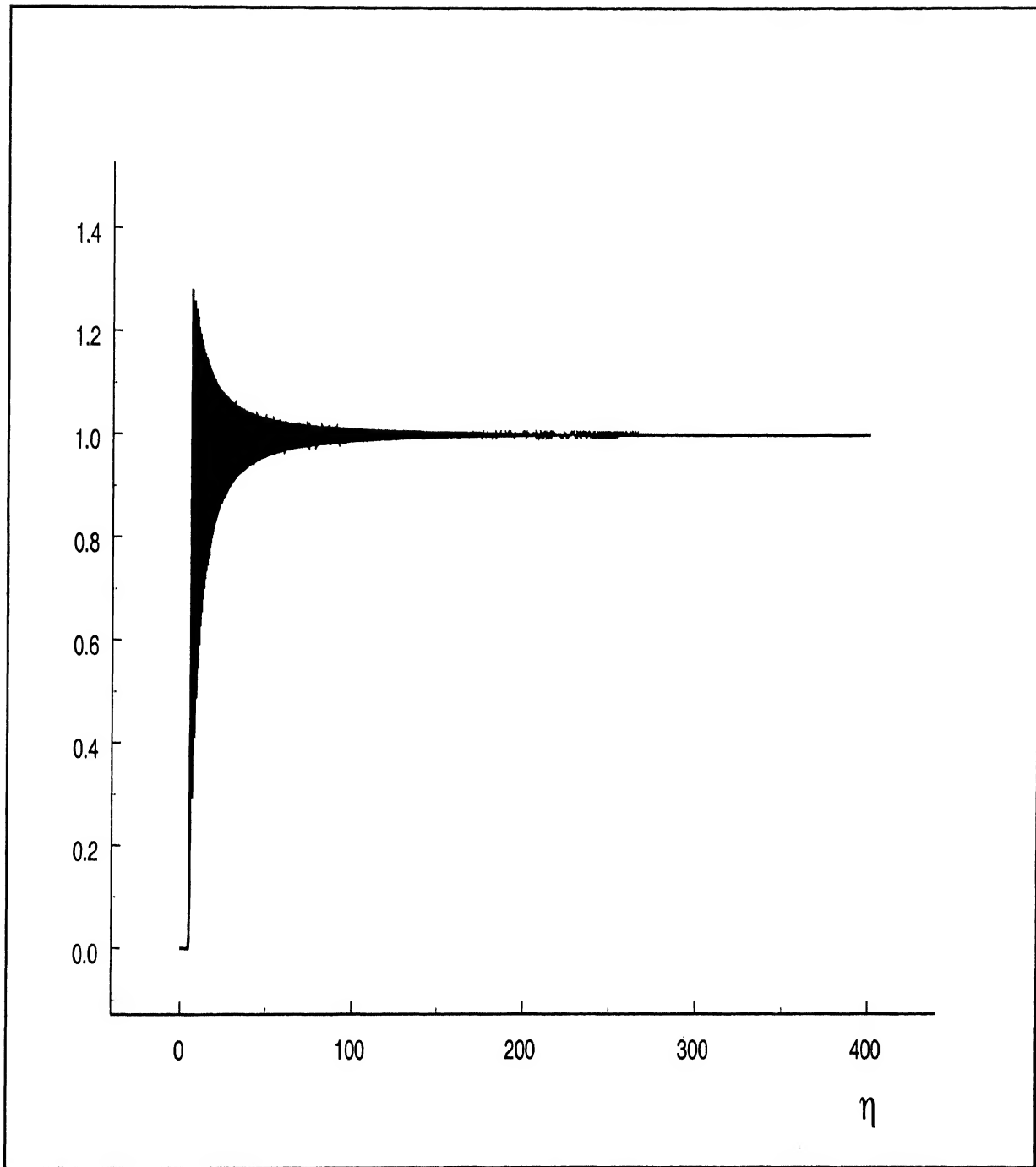


Figure 7:  $\frac{\lambda}{2Nm_0^2} \langle \vec{\Phi}^2 \rangle(\eta)$  vs.  $\eta$  (conformal time in units of  $m_0^{-1}$ ) for  $\frac{T_i}{T_c} = 3$ ,  $g = 10^{-5}$ . R.D. Universe.

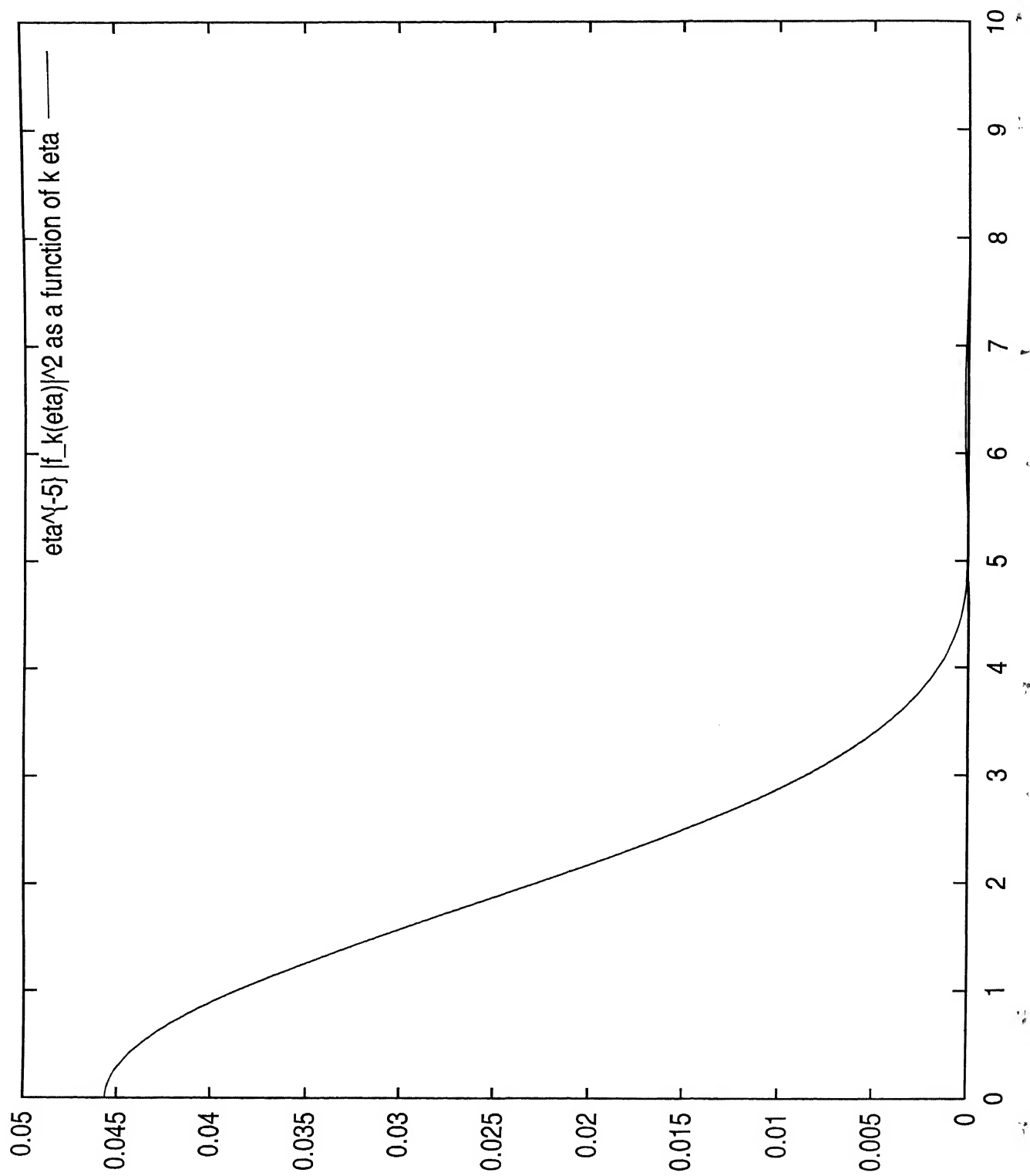


Figure 8:  $\eta^{-5}|f_k(\eta)|^2$  vs.  $k\eta$  for the same case as in Fig. 5.

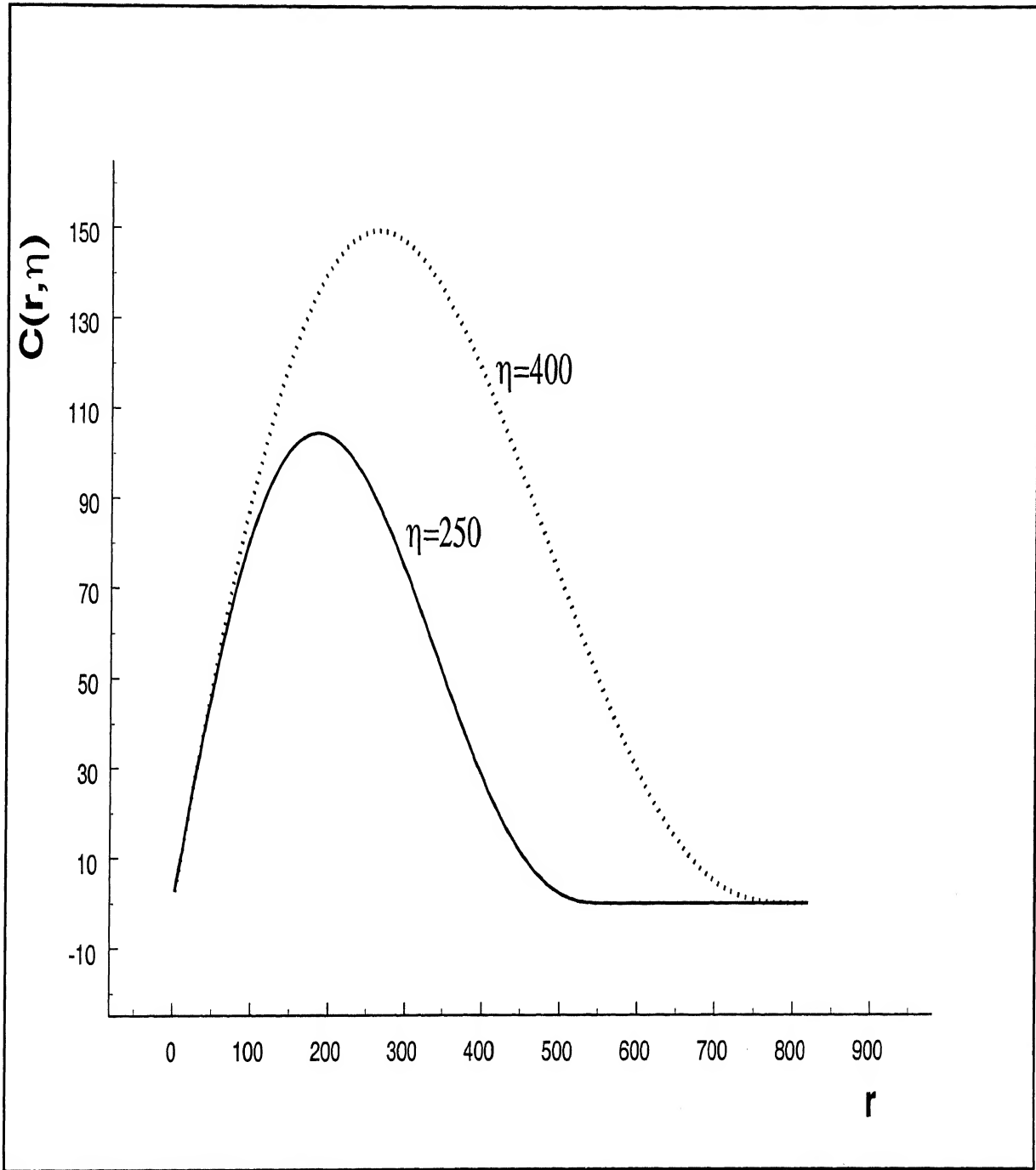


Figure 9:  $C(r, \eta)$  vs.  $r$  for  $\eta = 250, 400$  (in units of  $m_0^{-1}$ ) for  $\frac{T_i}{T_c} = 3$ ,  $g = 10^{-5}$ . R.D. FRW Universe.



# ORSAY LECTURES ON CONFINEMENT (I)

by

**Vladimir N. Gribov\***

L. D. Landau Institute for Theoretical Physics

Acad. of Sciences of the USSR, Leninsky pr. 53, 117 924 Moscow, Russia

and

KFKI Research Institute for Particle and Nuclear Physics

of the Hungarian Academy of Sciences

H – 1525 Budapest 114, P.O.B. 49, Hungary

and

Laboratoire de Physique Théorique et Hautes Energies\*\*

Université de Paris XI, bâtiment 211, 91405 Orsay Cedex, France

LPTHE Orsay 92/60

June 1993

---

\* Supported in part by Landau Institute-ENS Département de Physique exchange program

\*\* Laboratoire associé au Centre National de la Recherche Scientifique  
(Courtesy of Y. Dokshitzer, Ewarz and J. Nyiri)

## 1. The theory of supercharged nucleus

This talk is just introductory. Literally it is not about confinement, but it is important since it is closely connected to the theory of quark confinement. In fact, this theory deals with two different problems. The first one is, how to confine particles, i.e. the problem of producing forces strong enough to prevent quarks from separating, which will be discussed later. But there is also another, very severe problem.

We know, that quarks are light, almost massless particles. The question is, how to bind massless particles in a volume which is much smaller than the Compton wave length. Usually this is not easy to do, because the wave function is decreasing exponentially like  $e^{-\sqrt{m^2-\omega^2}r}$ , which in the best case can be  $e^{-mr}$ . But if the mass is very small, the state is very broad, and it is in complete contradiction with what we know about hadrons.

The mechanism of supercharged nuclei, which we will discuss in the present talk, seems to be a unique possibility to bind a particle in a small region in space. The theory of the supercharged nucleus is very old. It was initiated in the forties by the work of Pomeranchuk and Smorodinsky [1], and it became very well developed, with no unsolved questions left. Since we, however, want to apply this theory to quarks, we will talk about it in a way slightly different from what people are used to, combining the picture of the Dirac sea in external field and the language of Feynman's Green function.

The problem of the supercharged nucleus is the following. If there is a nucleus  $N_Z$  with a charge  $Z$ , and if  $Z$  is larger than a critical value  $Z_c$

$$Z > Z_c$$

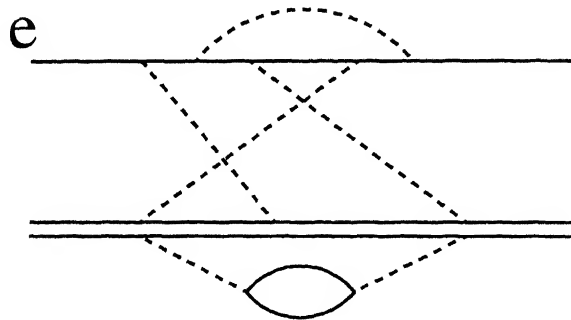
(theoretically  $Z_c = 137$ , practically it is around  $Z_c \sim 180$ ), then this nucleus will decay to an atom with a charge  $Z - 1$  and a positron:

$$N_Z \rightarrow A_{Z-1} + e^+ .$$

This atom can be stable or unstable. If it is unstable, it can decay again - up to a situation, when the total charge of the atomic state  $A_{Z-n}$  becomes sufficiently small. This is a peculiar thing : it means, that the nucleus behaves like a resonance. It does not exist in the nature freely, but it exists inside the atom. In this sense it is analogous to what we know about the quarks. This, of course, is not confinement, but in some respects it is not so different.

Indeed, the nucleus has a baryon number  $B$  and a lepton number zero. But this state with lepton number 0 does not exist, there are only states for which the lepton number equals unity. In this sense, it is a confinement of states with zero lepton number.

Theoretically the described problem is very well defined. If we want to understand the new type of atomic states, we have to consider the interaction of the electron with the nucleus, taking into account all possible interactions.



In this case there are some simplifications. First,

$$\alpha \ll 1 \quad , \quad (1)$$

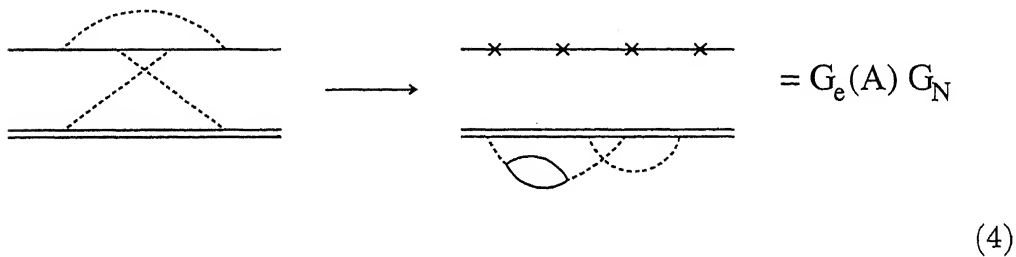
but

$$\alpha Z \sim 1 \quad . \quad (2)$$

The third important simplification is that the ratio of the electron mass  $m_e$  and the nucleon mass  $m_N$  is much smaller than unity :

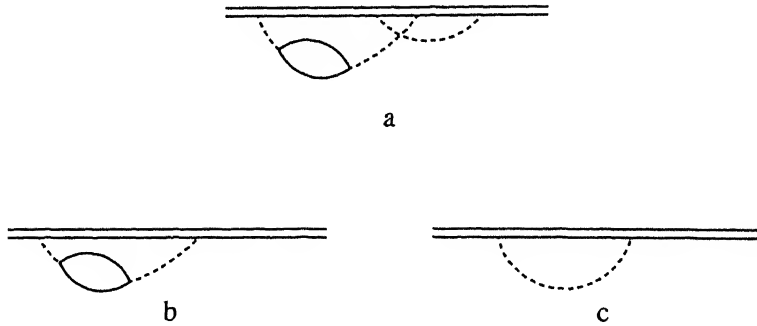
$$\frac{m_e}{m_N} \ll 1 \quad . \quad (3)$$

Because of these conditions, the problem can be solved exactly. First, due to (1), we can neglect all the corrections to electron propagation, since they would be of the order of  $\alpha$ , and leave those diagrams which contain lines connecting electron and nucleus lines and radiative correction to the nucleus line. But also, because of (3), the recoil of the nucleus due to photon emission will be very small, and instead of writing this interaction and taking into account all the sequences, including recoil, we can write that this is equal to the electronic Green function in the external field of the nucleus, multiplied by the nuclear Green function with radiative corrections





$$= G_e(A) G_N \quad (4)$$

Here  $G_e(A)$  is the Green function of the electron in the external field ; the Green function of the nucleus  $G_N$ , which corresponds to the diagram a



contains a lot of radiative corrections. We have to include loops like b since they are of the order of  $\alpha Z$ , and diagrams c being  $\alpha Z^2$ . Fortunately, all this turns out to be very simple. The reason is, that all the photons in these diagrams cause no recoil for the nucleus. It means, that we can write the following :

$$G_N = G_N^0 e^{\text{dashed circle} + \text{crossed circle}} \quad (5)$$

where everything can be factorized.  will be the symbolic expression for the electrostatic energy of the nucleus, and  gives us the mass renormalization for the nucleus due to electron vacuum polarization. And thus we have an exact description for  $G_N$ . Further, the first step is, of course, to calculate the electron Green function in an external field, which also defines the nucleus mass renormalization entering the exponent in (5).

This means, that with the knowledge of the Green function of the electron in the Coulomb field, we can calculate everything.

The equation for the electron Green function is

$$\widehat{\nabla} G_e = -i\delta \quad (6)$$

$$\gamma_\mu (\partial_\mu - iA_\mu) G_e = -i\delta \quad (7)$$

Since the potential does not depend on the time, we can always write

$$G_e(x_2, x_1) = \int \frac{d\omega}{2\pi i} e^{-i\omega(t_2-t_1)} \psi_\omega(r_1) \bar{\psi}_\omega(r_2) \quad (8)$$

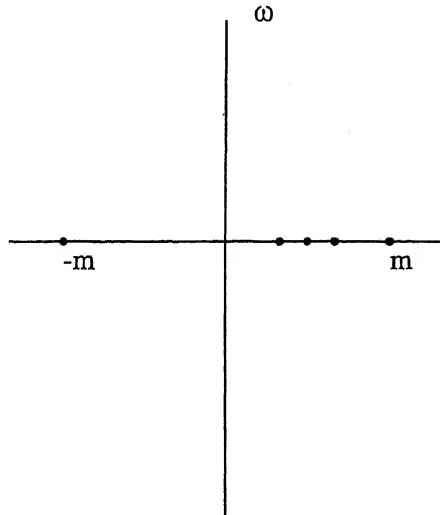
and, because of that, we have

$$\gamma_\mu (\partial_\mu - iA_\mu) \psi_\omega = 0 \quad , \quad (9)$$

or, in a more convenient way,

$$[\alpha_j p_j + m\gamma_0 - (\omega - A)] \psi_\omega = 0 \quad . \quad (10)$$

Before discussing the solution of the equation, let us consider the well-known features of the spectrum of this system.

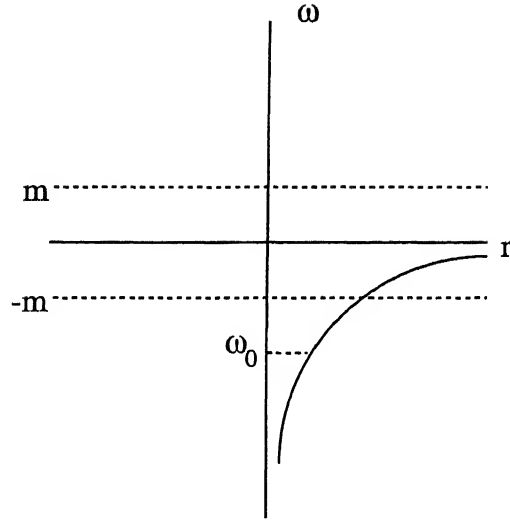


There are cuts going from  $m$  to  $\infty$  and from  $-m$  to  $-\infty$ , corresponding to the continuous spectrum, and sequences of poles corresponding to Coulomb levels.

Now - how does the critical phenomenon come in ? If we increase  $Z$ , the pole which corresponds to the atom in the ground state moves to the left, passing zero without any troubles (for a finite-size nucleus) and reaches the point  $-m$  at the value  $Z = Z_c$ . With the further growth of  $Z > Z_c$  the position of the level  $\omega_0$  is going to the complex plane and the state becomes unstable. However, at this point we come to a paradox : this  $\omega_0$  is supposed to be the energy of an atomic state, i.e. the energy of the atomic state becomes complex, contradicting the physics we have expected. Indeed, we thought that the nucleus was unstable and would decay on a stable atom and a positron.

This means, that the described simple solution has to be essentially changed, since, apparently, the Green function  $G_e(A)$  of the electron does not reflect the whole physics, and the features of  $G_N$  in (4) turn out to be important. We will show, in fact, that there is a cancellation between  $G_e(A)$  and  $G_N$ .

Let us consider now, what is happening from the point of view of the Dirac sea. From  $m$  to  $\infty$  and from  $-m$  to  $-\infty$  there is a continuous spectrum of electron states in the Coulomb field. Also, as we said before, there is a resonance state with complex energy  $\omega_0 + i\Gamma/2$ , where  $\omega_0 < -m$ . What is the physics of this resonance ?



The Dirac equation can be written in the form

$$(T + A)\psi = 0 \quad (11)$$

where  $T$  is the kinetic energy and  $A$  the potential energy. This kinetic energy can have two signs  $T = \pm\sqrt{p^2 + m^2}$ , and therefore the total energy can be quasi-classically either

$$\sqrt{p^2 + m^2} + A(r) = E \quad (12)$$

or

$$-\sqrt{p^2 + m^2} + A(r) = E \quad (13)$$

Obviously, (12) can be fulfilled only close to the origin. In this case the potential is negative, the square root is positive, and the total energy is negative. Classically, the electron will be stopped at a return point  $r_1$ , where the energy  $E$  is the sum of the potential and the mass  $m$  :

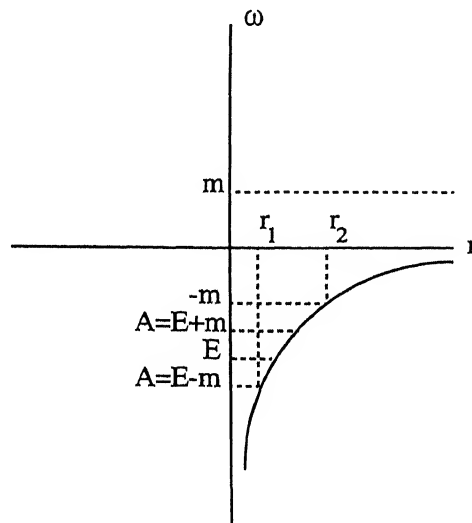
$$m + A(r_1) = E \quad (14)$$



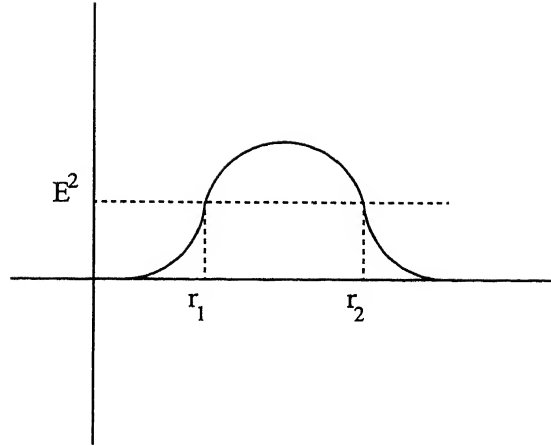
This means, that in this region there is a normal wave function which oscillates at  $r < r_1$ . After that it has to decrease exponentially because of the absence of classical trajectories. For the equation (13) the situation is the opposite, it can exist only at large distances, since the potential at large distances is small, and the energy is negative ; consequently, there will be another return point  $r_2$ , where

$$-m + A(r_2) = E \quad (15)$$

Again, there will be a plane wave, i.e. normal oscillation at  $r > r_2$  with a subsequent exponential decrease due to the absence of classical motion between the points  $r_1$  and  $r_2$



The situation is somewhat similar to that with a barrier in the non-relativistic problem :



This barrier-type behaviour can be easily seen, if one writes a second order equation instead of the first order one. Indeed, let us re-write the Dirac equation in the form

$$\left( \nabla^2 + \frac{1}{2} F_{\mu\nu} \sigma_{\mu\nu} \right) \psi = 0 . \quad (16)$$

Here

$$\nabla^2 - (\partial_t^2 - iA)^2 = -\frac{1}{r} \partial_r^2 r - \omega^2 + 2\omega A - A^2 . \quad (17)$$

In this equation the effective potential energy is  $2\omega A - A^2$ . If  $A$  is negative, the first term  $2\omega A$  corresponds to attraction at large distances, and the second term  $-A^2$  to repulsion at short distances.

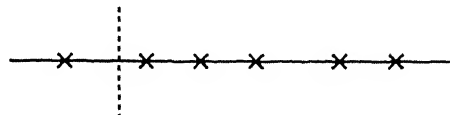
It is natural to expect, that this negative energy state is just a resonance. But : in the previous discussion we forgot about the Pauli principle. We have been talking only about one state and not about the Dirac sea. However, Dirac sea means, that all these normal negative levels have to be occupied.

In this case, if our particle will try to go out and pass through this barrier, there will be no place for it. When we find, that the Green function  $G_e$  has a complex singularity, this reflects the fact, that we did not say up to now, what type of Green function we are using.

Imagine, that the Dirac sea is not filled up. In this case it would be natural for our state to have complex energy. On the other hand, we are used to the fact, that the Feynman Green function reflects the Pauli principle in the correct way, and therefore, if we are looking for the Feynman Green function, we expect to find the proper answer.

Suppose now, that we will calculate the Feynman Green function. Still, we said that there is a general solution for the Green function with a complex pole. The question is, how this happens.

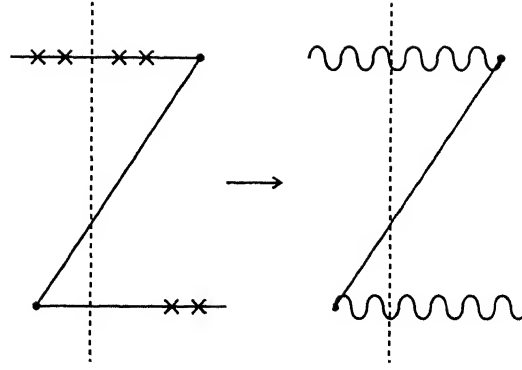
The answer is essentially very simple. Indeed, we are looking for a Green function which is a sum of diagrams of the type


(18)

If the Feynman Green function contained only positive energy propagating in positive time, all the energy denominators in (18)

$$\frac{1}{\varepsilon - \sqrt{m^2 - p^2}}$$

would be real at negative total energy. But the Feynman Green function contains also negative energy, propagating in negative time. It corresponds to the so-called  $Z$ -diagram

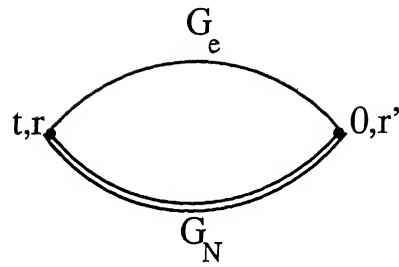


Perturbatively this diagram contains also only positive energy intermediate states. However, if the incoming and outgoing lines will correspond to negative energy bound states, then the three-particle intermediate state containing two negative energy particles and one positive energy particle can have negative total energy :

$$\frac{1}{\varepsilon - \sqrt{m^2 + p^2} - 2\varepsilon} = -\frac{1}{\varepsilon + \sqrt{m^2 + p^2}}$$

This means, that the Feynman Green function in this case obtains an imaginary position of the pole in contradiction with the Pauli principle. (This contradiction is obvious from the second of the  $Z$ -diagrams, since we have there two particles in the same state at the same time). It is also clear, that this diagram reflects not the decay of a state, but that of the vacuum.

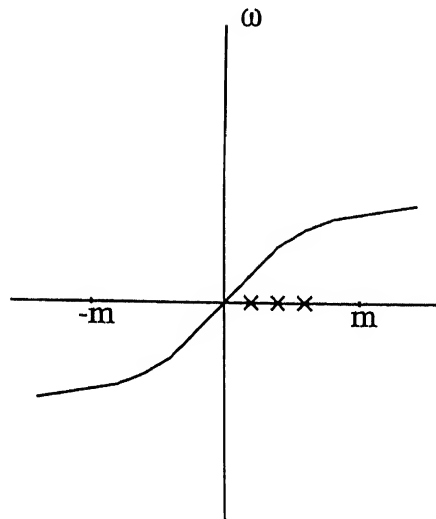
Let us see, how the Green function of the nucleus will recover the Pauli principle. Consider the diagram



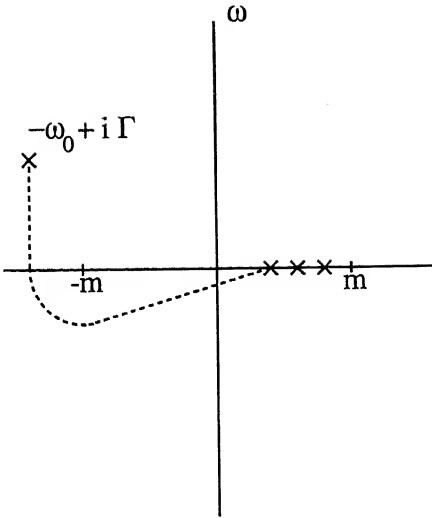
The Green function of the electron is in this case

$$G_e(t, r, r') = \int \frac{d\omega}{2\pi i} e^{-i\omega t} \psi_\omega(r) \bar{\psi}_\omega(r') . \quad (19)$$

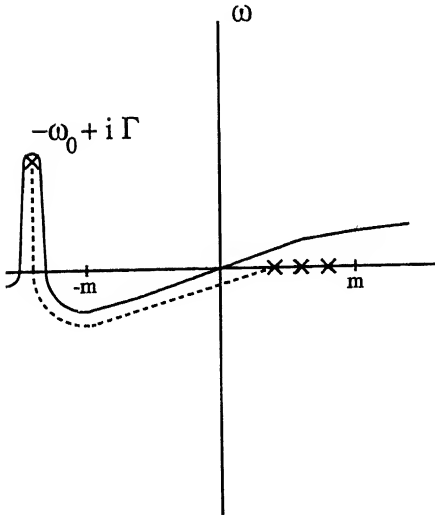
In order to calculate it, we have to write the following contour of integration (for  $Z < Z_c$ )



With  $Z$  growing, the pole will move (as indicated by the dotted line) to the complex plane,



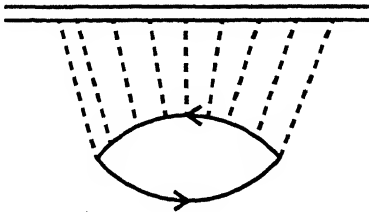
and we will have to change the contour of integration:



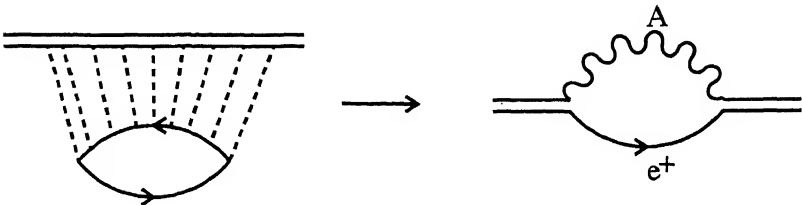
As we see, the Green function will contain a growing exponent which will be defined by the imaginary part of the pole. At large  $t$  we will be left with an unstable pole contribution

$$G_e(t, r, r') = \int \frac{d\omega}{2\pi i} e^{-i\omega t} \psi_\omega(r) \bar{\psi}_\omega(r') \sim e^{\Gamma t - i\omega_0 t} .$$

Obviously, the total of the diagrams can not contradict the Pauli principle, and due to this fact there has to be a cancellation between the electron Green function and the Green function of the nucleus. The latter contains contributions of the type



and has singularities connected with atomic states

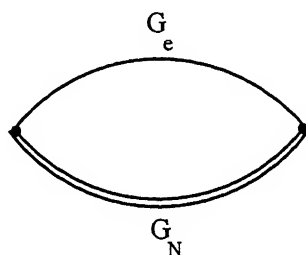


This means, that the Green function of the nucleus will contain this loop, which, definitely, has an imaginary part, because the nucleus becomes heavier than the atom and the positron. It gives us a contribution which describes

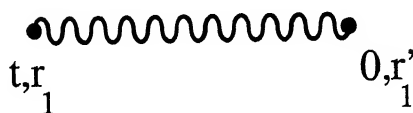
the instability and results in an exponent  $e^{-\Gamma t}$  corresponding to the decay of the nucleus

$$\text{diagram} + \text{diagram} + \dots = e^{\text{diagram}} = e^{-\Gamma t - i M t}$$

Hence,  $e^{\Gamma t}$  will cancel and

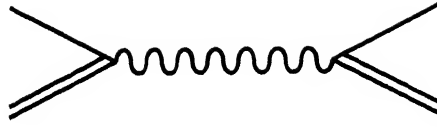


will be proportional to  $e^{-i(\omega_0 + M)t}$  which corresponds to the propagation of an atomic state with real energy  $\omega_0 + M$

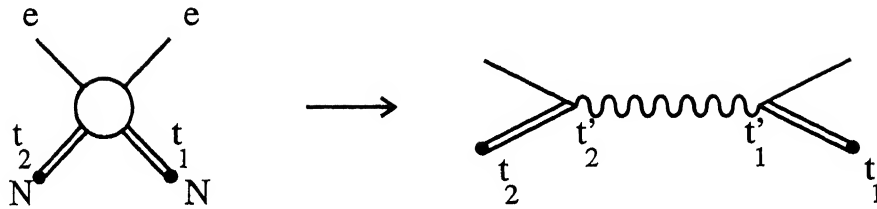


So, indeed, we have shown, that in this case the bound state is stable not only due to the existence of binding forces, but also because of the Pauli principle, which reflects the antisymmetry between our electron and the electron in the vacuum. In the electron-nucleus scattering amplitude (4) asymptotically (i.e. at large  $t$ ) only the part





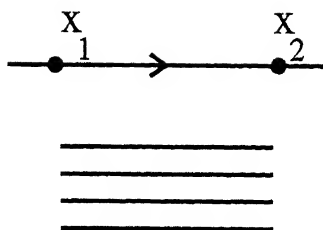
survives in the sense



because the Green function of the nucleus decays always exponentially, and it has to be compensated by the increase of the electron Green function. This is possible only, if

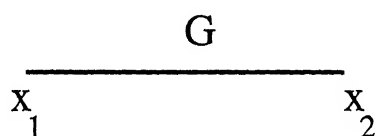
$$\Gamma(t_2 - t'_1) < 1, \Gamma(t'_1 - t_1) < 1, t'_2 - t'_1 \rightarrow \infty.$$

Let us see now, how we have lost the Pauli principle in discussing Feynman Green functions. This can be understood immediately. Our aim is to derive the Feynman prescription from the Dirac picture. In the latter we have negative levels filled up, and we look for the propagation of additional particles.

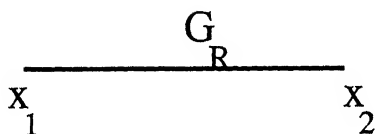


How can we derive Feynman rules from this ? It will be, indeed, very simple.

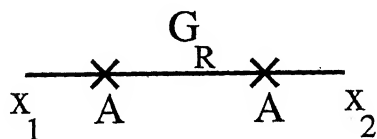
In this picture our Green function



has, of course, to be retarded :

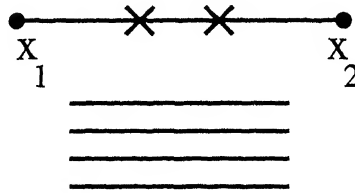


The real propagation is second order in external field :

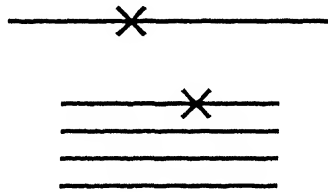


(20)

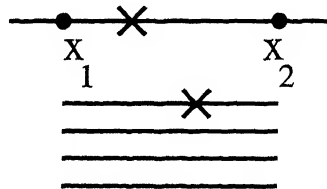
We start with the usual calculation, but we have to include the Dirac sea. The external field  $A$  is acting not only on the particles, but also on the Dirac sea. This means, that we have to consider our particle and a particle from the Dirac sea ;



will be a possible diagram. However, we can write also

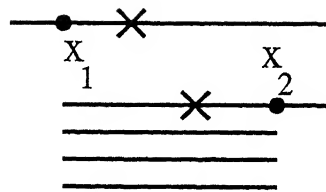


and consider the diagrams



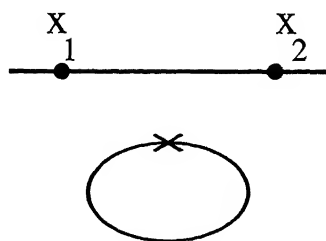
(21)

and

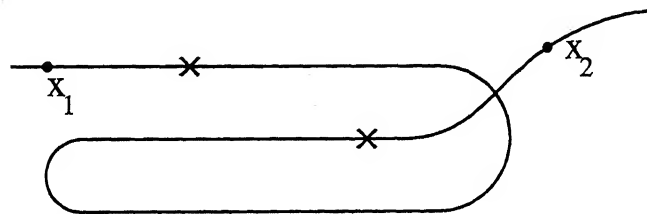


(22)

Averaging over all particles in the Dirac sea, the diagram (21) leads to



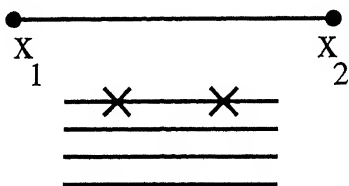
which corresponds to the first order contribution to vacuum polarization and is in fact zero. The diagram (22) gives



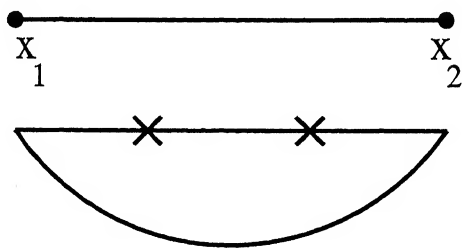
and we get

$$\begin{array}{c}
 \times \quad \times \\
 \hline
 x_1 \quad G_- \quad x_2
 \end{array} \quad (23)$$

summing all negative energy levels. Adding the diagram with  $G_R$  (20) to that with  $G_-$  (23), we obtain the Feynman diagram. By these two diagrams, however, symmetry is not reflected. In order to be symmetrical, we have to add



which means that the external field is interacting twice with the vacuum. But averaging this, we get



i.e. we have to add a diagram which corresponds to vacuum polarization in external field. In other words, the Pauli principle is taken into account by adding vacuum polarization to the normal Feynman propagator.

It is, of course, always necessary to consider vacuum polarization, but sometimes it is relevant, sometimes not. In the case of weak coupling the vacuum polarization was real and, when added to the Feynman propagator, did not change the picture essentially. But in our case, in the case of strong coupling, the vacuum polarization becomes complex and because of this, it defines everything.

In conclusion, let us make it clear, which region of distances between the nucleus and the electron is important in our calculation.

As it is shown in [2], the solution  $\psi_\omega(r)$  of the Dirac equation (10) is

$$\psi_\omega(r) \sim \frac{1}{\sqrt{r}} \cos \left[ \sqrt{(Z\alpha)^2 - 1} \ln \frac{r}{r_0} + \delta \right] \quad (24)$$

in the region

$$r_0 \ll r \ll \left| \frac{1}{\omega} \right| ,$$

where  $r_0$  is the radius of the nucleus. If  $|\omega|r_0 \ll 1$ , the function  $\psi_\omega(r)$  will oscillate and at  $r_0 \rightarrow 0$  this corresponds to "falling into the center". If  $r_0$  is finite, the "falling into the center" does not occur. However, the existence of oscillations is a sign which indicates that the levels passed through the point  $\omega = -m$ .

The number of oscillations  $n$  in the region  $r_0 < r < \frac{1}{m}$  which determines the number of levels passing through is defined by the condition

$$\sqrt{(Z\alpha)^2 - 1} \ln \frac{1}{mr_0} + \delta = n\pi . \quad (25)$$

At  $n = 1$  formula (25) provides the condition for the charge to be supercritical (with  $r_0$  and  $m$  finite) and shows, that the region where the supercritical phenomenon occurs is

$$r_0 \ll r \leq \frac{1}{m} .$$

### Acknowledgements

The text of these lectures was written by Julia Nyiri. She substantially expanded and edited the original notes she took during the lectures. I am very grateful to Ph. Boucaud, A. Dudas and J. Mourad who formatted the text and the figures. Also, I would like to thank the Lab. de Physique Theorique et Hautes Energies, and especially A. Capella, M. Fontannaz, A. Krzywicki, D. Schiff and Tran Than Van for their hospitality, and the warm and inspiring atmosphere during my stay at Orsay.

### References

- [1] I. Pomeranchuk, Ya. Smorodinsky, Journ. Fiz. USSR 9 (1945) 97.
- [2] Ya. Zeldovich, V. Popov, Uspekhi Fiz. Nauk 105 (1971) 4.

# Orsay Lectures On Confinement (II)

Vladimir N. Gribov\*

L. D. Landau Institute for Theoretical Physics Acad. of Sciences of the USSR  
, Leninsky pr. 53, 117 924 Moscow, Russia

and

KFKI Research Institute for Particle and Nuclear Physics  
of the Hungarian Academy of Sciences,  
H-1525 Budapest 114, P.O.B. 49, Hungary

and

Laboratoire de Physique Théorique et Hautes Energies<sup>†</sup>  
Université de Paris XI, bâtiment 211, 91405 Orsay Cedex, France

LPTHE Orsay 94-20  
February 1994

---

\*Supported in part by Landau Institute-ENS Département de Physique exchange program

<sup>†</sup>Laboratoire associé au Centre National de la Recherche Scientifique



## 1 The confinement of the heavy quark

In the previous talk [1] we have considered some aspects of the theory of supercharged nuclei, and came to the conclusion, that the superbound atoms are stable mainly due to the Pauli principle.

Before going to the heavy quark, let us discuss briefly, how the problem of the supercharged nucleus can be handled practically. One possibility to formalize it is the following. We have to calculate the Green function of the Dirac equation in external field :

$$\hat{\nabla} G(x, x') = \frac{1}{i} \delta(x - x') \quad , \quad (1)$$

where

$$\hat{\nabla} = \gamma_\mu (\partial_\mu - iA_\mu) \quad . \quad (2)$$

The initial external field is that of the considered nucleus. However, in the presence of a charge  $Z$  the stationary states correspond to atomic states with a charge  $Z - N$  where  $N$  is large enough to fulfil  $Z - N < Z_{cr}$ , i.e. there have to be  $N$  electrons rotating around the nucleus. But if so, we will have to take into account, that the field which acts on each electron is also changing.

The ground state in our case is an atomic type state. This means, that although we don't know what our field is, we can find it. The gradient squared of the zeroth component  $A_0 \equiv A$  of the potential has to be equal

$$\nabla^2 A_0 = e_0^2 (\rho_{nuc} + \rho_{el}) \quad . \quad (3)$$

where  $\rho_{el} = \langle \bar{\psi} \gamma_0 \psi \rangle$ .

The charge density will depend on the potential, and the equation becomes a Thomas-Fermi type equation for the potential existing in this system. If we find a wave function satisfying the equation we will know  $A$ , the Green function and  $\rho$ . This means, that the formal way of solution is just to solve a self-consistent, Thomas-Fermi type equation for the effective atomic potential.

Let us consider the total charge  $Q$

$$Q = \int \rho d^3r \quad ; \quad (4)$$

the density is the average value

$$\rho = \langle \bar{\psi}(r) \psi(r) \rangle = \sum_{\omega < 0} \bar{\psi}_n(r) \psi_n(r) \quad (5)$$

which in the Dirac picture is the sum of all negative energy levels. The total charge  $Q$  will be the sum of the energy levels over all  $\omega_n$ . The sum is divergent, and we have to make a cut-off and subtract the bare particles. This, however, will not be enough. Indeed, if we just subtract the value of the charge of the free vacuum, then at  $Z \neq 0$  the charge of the vacuum (i.e. of the nucleus) will not be equal  $Z$  and will continue to diverge logarithmically. We have to subtract the value at small  $Z$  in a way which gives zero for the total charge of the electron vacuum (this corresponds to the correct renormalization of the charge of the nucleus).

The above procedure does not reflect literally the subtraction of the vacuum charge without external field. This becomes especially obvious, if one makes use of the Levinson theorem which connects the number of states of a particle in external field in a given energy interval with the phase of scattering of this particle in the same field. The number  $N$  of additional states in the energy interval between  $E_1$  and  $E_2$  is defined by the difference of phase shift

$$N = \frac{1}{\pi} [\delta(E_2) - \delta(E_1)] \quad (6)$$

Because of this, the number of new states in the Dirac vacuum equals

$$N = \frac{1}{\pi} [\delta(-m) - \delta(-\infty)] \quad (7)$$

For  $Z < Z_c$  this number should be zero. However, for the Dirac equation in external field this condition is not fulfilled, because  $\delta(-\infty) \neq 0$ , and in order to obtain the proper definition of the vacuum charge we have to subtract a quantity which, generally speaking, depends on  $Z$  ( $\delta(-\infty)$ ). This subtraction means, in fact, that we have to change the interaction with the external field when  $E \rightarrow -\infty$  so that  $\delta(-\infty) = 0$ . ( $\delta(-m)$  can always be considered to be zero if the field is small). If, however,  $Z$  will be increased and becomes more than critical, then, as we saw, the levels will pass the point  $E = -m$  and move to the complex plane. It is easy to check, that every time this happens the phase  $\delta(-m)$  is changing by  $\pi$  and the move of  $n$  levels into the complex plane changes the number of states by  $n$  so that the charge of the vacuum (i.e. of the atom) becomes  $Z - n$ . The value of  $\rho_e$  which enters the equation for the self-consistent field is to be defined by the contribution of levels which passed through  $E = -m$ .

We discussed in detail this concrete structure in order to refer to it in the following, talking about the heavy quark in the vacuum of the light quark. We shall suppose, that due to gluonic vacuum polarization the effective coupling  $\alpha(r)$  which at small  $r$  has the usual perturbative behaviour, reaches a constant value at  $r \gg r_0$  ( $r_0 \sim 1/\lambda$ ), and this constant value will be more than unity. (Without this ansatz, allowing that  $\alpha$  continues to grow, things are more complicated, but nothing essentially changes).

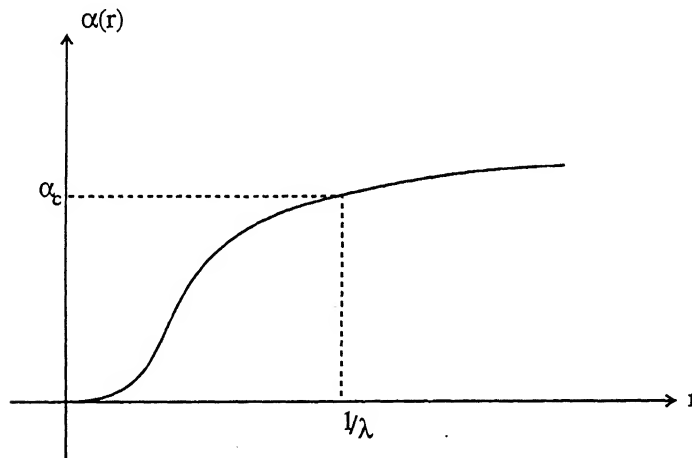
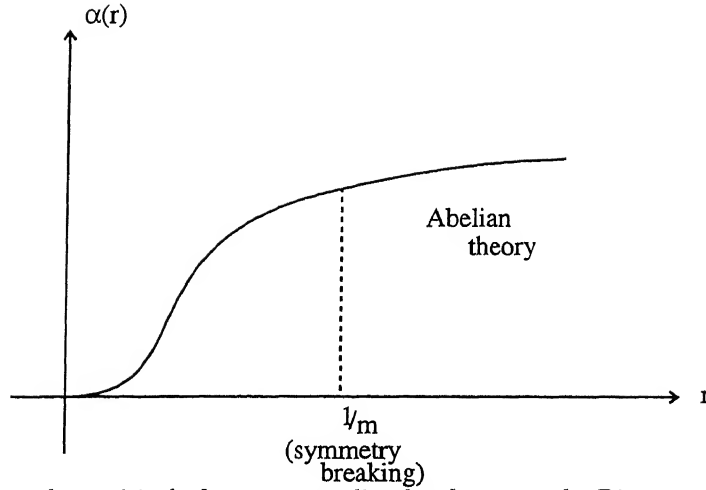


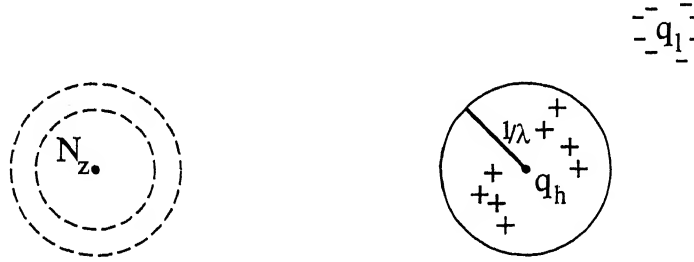
Figure 1

The supposed behaviour of the effective coupling.

It is natural to expect such a behaviour of the charge in QCD. But even an Abelian theory can reveal such a behaviour of charge, if it originates from a non-Abelian theory via spontaneous colour symmetry breaking. In this case the charge will increase, and as a result, 6-7 gluons acquire masses. After that there remain one or two massless objects - "photons" - and the behaviour of charge will be exactly as discussed, because after the symmetry breaking we have  $r_0 = \frac{1}{m}$  (heavy gluon). We can ask now : what happens in the vacuum of light quarks under these circumstances ? Outside the region where  $\alpha$  is growing, we will have an Abelian theory and we can consider the quark states in the normal way.



If  $\alpha$  becomes more than critical, the corresponding level goes to the Dirac sea, which, consequently, will have to be filled up and we will find an atomic type state. If this is so, the charge density will be exactly the same as in the case of a supercharged nucleus. The  $\lambda_{QCD}$  is analogous to the radius of the nucleus. There is, however, an important difference between QED and this case. Indeed, in QED we consider a nucleus with a charge  $Z$  in the centre, and we put one or two electrons around it to organize an atom. In QCD we have a heavy quark with a very small intrinsic charge.



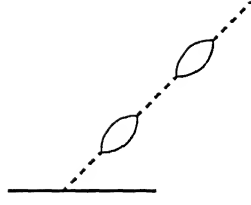
Due to vacuum polarization the charge becomes large inside  $1/\lambda$ . The system creates an empty level, and we have to fill it. This means, that we have to add an intrinsic charge, equal to that of the heavy quark, but with the opposite sign (outside the region of  $1/\lambda$ ). So we have here two intrinsic charges with the total charge zero, and of course the vacuum polarization can never change the total charge. In other words, in QCD the supercritical atom would be a meson-type state with zero colour charge. Our task is to show, that these are not only words - we have to include the mechanism of vacuum polarization formally. In order to do so, let us discuss the problem in a language similar to QED, neglecting non-Abelian fluctuations of the colour field  $A$  - the average field of heavy and light quarks inside the meson. In this case  $A_0$  is defined by the equation

$$\nabla^2 A = e_0^2 (\rho_h - \rho_l) \quad , \quad (8)$$

where  $\rho_h$  is the density of the heavy quark, and  $\rho_l$  that of the light quark. Again, we can write the equation for the Green function

$$(\hat{\nabla} - m) G(x, x') = \frac{1}{i} \delta(x) \quad . \quad (9)$$

But now the problem is, how to calculate this. What means, e. g.,  $\rho_{heavy}$  ? We can not use the expression  $\rho_h = e_0^2 \delta(r)$ , since it does not include the bosonic vacuum polarization, which has to be taken into account. It is well known, that in principle vacuum polarization means summing a diagram like



with gluonic loops inside. The problem is, however, how to write this in normal space-time language. There is a good way to sum this diagram by writing an equation for currents. Knowing the external current and wishing to calculate the total current, we have to solve the equation

$$j_\mu(x) = j_\mu^{ext}(x) + \frac{\alpha_0 b}{2\pi^2} \int \delta'((x-x')^2) j_\mu(x') d^4 x' , \quad (10)$$

which corresponds to the summation of the diagram. We will not derive this equation, because it is almost obvious for the static charge which we are interested in.

Let us consider a static charge, not depending on time. In this case

$$\rho(r) = \rho_\mu^{ext}(r) + \frac{\alpha_0 b}{8\pi^2} \int \frac{d^3 r'}{|r-r'|^3} \rho(r') , \quad |r-r'| > \varepsilon , \quad (11)$$

where  $1/\varepsilon$  is the ultraviolet cut-off,  $b = \frac{11}{3}n_c - \frac{2}{3}n_f$ .

The solution of this equation is very simple and leads to the usual expression for charge renormalization in QCD. In order to see this, let us introduce the quantity  $Q(r)$  :

$$Q(r) = \int_0^r \rho(r') 4\pi r'^2 dr' . \quad (12)$$

The logarithmic derivative of  $Q(r)$  is

$$\partial_\xi Q(r) = r \frac{\partial Q(r)}{\partial r} = 4\pi r^3 \rho(r) . \quad (13)$$

For  $Q(r)$  we can write the equation

$$\partial_\xi Q(r) = \partial_\xi Q_{ext}(r) + \frac{\alpha_0 b}{8\pi^2} \int_{|r-r'| > \varepsilon} \frac{d^3 r'}{|r-r'|^3} \frac{r^3}{r'^3} \partial_{\xi'} Q(r') . \quad (14)$$

The integration in the right hand-side of this expression contains two logarithmic regions :  $|r-r'| < r$  and  $r \gg r'$ . In the first region,  $\partial_{\xi'} Q(r')$  can be substituted by  $\partial_\xi Q(r)$  ; integrating over the second region,  $r'$  in the denominator  $|r-r'|^3$  can be neglected. As a result, we obtain

$$\partial_\xi Q(r) = \partial_\xi Q_{ext}(r) + \frac{\alpha_0 b}{8\pi^2} \int_\varepsilon^r \frac{d^3 r''}{r''^3} \partial_\xi Q(r) + \frac{\alpha_0 b}{2\pi} Q(r) ,$$

which is equivalent to

$$\partial_\xi \left( 1 - \frac{\alpha_0 b}{2\pi} \ell n \frac{r}{\varepsilon} \right) Q(r) = \partial_\xi Q_{ext}(r) \quad (15)$$

or

$$Q(r) = \frac{Q_{ext}(r)}{1 - \frac{\alpha_0 b}{2\pi} \ell n \frac{r}{\varepsilon}} . \quad (16)$$

For a point-like charge

$$Q_{ext}(r) = 1 ,$$

and we have

$$A(r) = \frac{\alpha(r)}{r} = \alpha(r) A_{ext} \quad . \quad (17)$$

The concrete expression

$$\alpha(r) = \frac{\alpha_0}{1 - \frac{\alpha_0 b}{2\pi} \ell n \frac{r}{\epsilon}} \quad (18)$$

obtained from perturbation theory has, of course, an unphysical singularity. For a point-like charge (18) has to be substituted by an expression corresponding to the behaviour of  $\alpha(r)$  as shown in Fig. 1. However, for the distributed charge the relation between the external field and the field which takes into account the polarization is non-local in coordinate space. The correct expression for the relation between the external field and the observable field is local in the momentum space :

$$A(q) = \alpha(q^2) A_{ext}(q) \quad (19)$$

which leads in the coordinate space to an expression of the following type :

$$A(r) = \int K(r - r') A_{ext}(r') d^3 r' \quad , \quad (20)$$

where

$$K(r) = \int e^{iqr} \alpha(q) \frac{d^3 q}{(2\pi)^3} \quad .$$

Similarly,

$$\rho(r) = \int K(r - r') \rho_{ext}(r') d^3 r' \quad . \quad (21)$$

If we now suppose, that  $\alpha(q)$  as a function of  $1/q$  behaves according to Fig. 1, and at large  $q$  values it is defined by perturbation theory, then the equation (8) has to be understood as an equation for  $A_{ext}$  :

$$\nabla^2 A_{ext}(r) = \delta(r) - \bar{\psi}_\ell(r) \gamma_0 \psi_\ell(r) \quad (22)$$

where  $\psi_\ell(r)$  is the solution of the Dirac equation

$$(H + A)\psi = E\psi$$

for the light antiquark in a superbound state in the field  $A(r)$  defined by equation (20). The solution of this problem gives the energy and the features of the meson  $q_n \bar{q}_\ell$  with zero total charge. Due to (21),

$$Q = \int K(r) d^3 r \quad Q_{ext} = \alpha(q=0) Q_{ext} \quad (23)$$

with  $Q = 0$  if  $Q_{ext} = 0$ .

We have just proved, that because of the big charge which appears through vacuum polarization, in the case of QCD the “atomic” bound state will, indeed, be a meson. The heavy quark will decay into a  $q_h \bar{q}_\ell$ -meson and a light quark :

$$q_h \rightarrow M(q_h \bar{q}_\ell) + q_\ell \quad . \quad (24)$$

In the next lecture we shall consider light quark states. So far there is one important thing to stress, namely if we don't include essential interactions between light quarks in the vacuum, we come to a reasonable conclusion for the case of the heavy quark but, as we will just see, to an unreasonable one about the light quark. Indeed, let us try to extend the considered procedure to

the latter case. Suppose, that there is a light quark moving, and a potential acts on its vacuum. What will we see classically ? Since the Coulomb field is a vector field, it is shrinking, but the total integral remains the same. Because of this, we will find immediately, that there is a bound state in this potential even for fastly moving particles. This, however, means, that we have here an unstable state, which has to be filled, and as a consequence the light quark will decay into a meson and a light quark again :

$$q\ell \rightarrow M + q\ell$$

which, of course, contradicts the energy conservation, unless the appearing meson is of negative energy. In order to have a self-consistent picture, we have to suppose that the light quark in the vacuum will interact so strongly that there have to be negative energy levels and the whole vacuum has to be rearranged. So from the picture we described we come quite naturally to light quark interactions. We will see, that these interactions are, indeed, very strong and lead to the confinement of light quarks which will take place at relatively small  $\frac{\alpha}{\pi}$  values ; this means, that the overall corrections for vacuum polarization will not be large.

This is for the future. What we have to add now, is, that even in the language which was accepted so far, with no strong interactions between particles in the vacuum, the problem in real QCD which is non-Abelian is more complicated. In QED we have one charge, and all the electrons in the vacuum interact with this charge, independently from each other. In QCD this can take place only, if the field of the heavy quark is an Abelian one. It is highly probable, that this is, indeed, the case, when the field of the heavy quark becomes large as a result of gluonic vacuum polarization.

## References

- [1] V.N. Gribov, Orsay Lectures on confinement (I): The theory of supercharged nucleus, LPTHE ORSAY 92/60 (1993).

# Orsay Lectures On Confinement (III)

V. N. Gribov

L.D. Landau Institute for Theoretical Physics, Moscow

and

KFKI Research Institute for Particle and Nuclear Physics, Budapest

and

Laboratoire de Physique Théorique et Haute Energies

Université de Paris XI, bâtiment 211, 91405 Orsay Cedex, France

LPT-ORSAY-99-37  
hep-ph/9905285

## Light quark confinement<sup>1 2</sup>

We have described the confinement of heavy quarks in an analogy with the theory of the super-charged nucleus [1,2]. Let us now suppose again that  $\alpha$  is behaving like

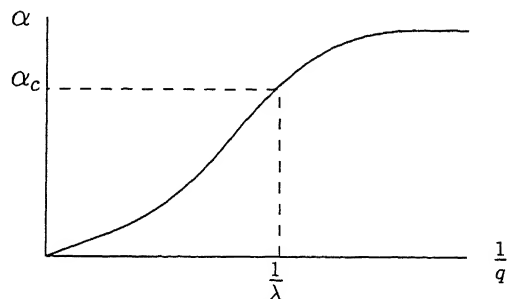
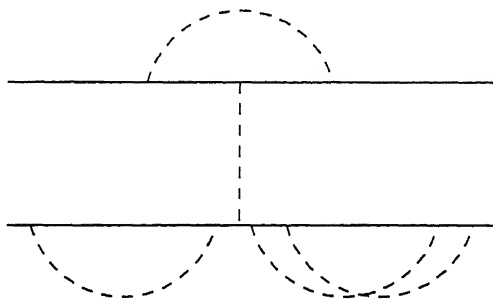


Fig. 1.

Making this assumption, we are neglecting gluon-gluon interactions and the existence of gluons as real particles. Our aim is to see, what can arise from the discussion of light quarks only. We introduce  $\lambda$  corresponding to  $\alpha_c$  and consider quark masses  $m_0 \ll \lambda$ . The interactions of light quarks (for which  $m_0 \ll \lambda$ ) will be discussed in a rather simplified way. We will take into account all possible interactions

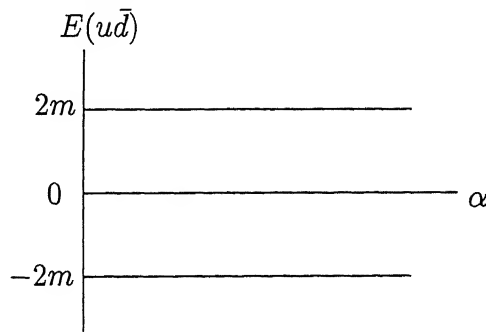


where the gluon propagator (considered as an effective photon), corresponding to the dotted line is

$$D_{\mu\nu} = \frac{\alpha}{q^2} \delta_{\mu\nu}. \quad (1)$$

Further, we look for a model which enables us to see, what happens to the fermions if there is an interaction between them, as indicated above. The question is, how the bound states or the Green function behave in such a case.

Let us consider the energy of two quarks,  $u$  and  $\bar{d}$ , for example. Without interaction there will be positive energy states with  $E > 2m$  and negative energy states with  $E < -2m$ :

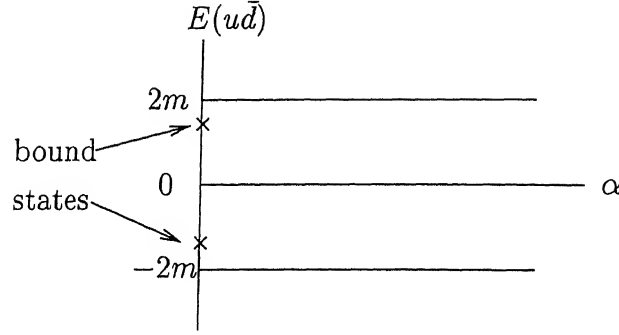


<sup>1</sup>This is the third lecture on quark confinement given by V.N.Gribov in 1992 in Orsay. An extensive discussion of the consequences of all this for the structure of the Green function can be found in [5,6] - in the two last papers concluding his 20 years long study of the problem of quark confinement in QCD.

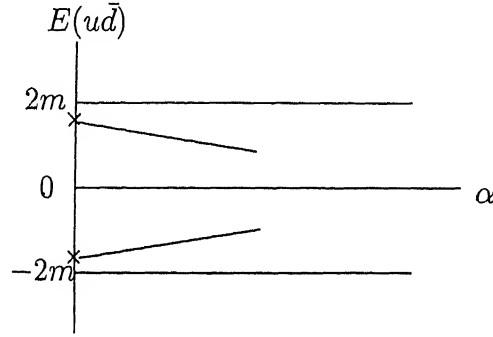
<sup>2</sup>The text was prepared for publication by Yu. Dokshitzer, B. Metsch and J. Nyiri on the basis of a tape recording and notes taken during the lecture



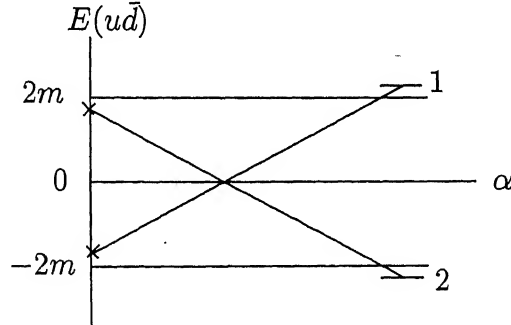
Introducing the interaction, for small  $\alpha$  we will find that there are some bound states near  $2m$  and  $-2m$ .



So far, we consider the usual Dirac vacuum: the negative energy states are occupied, and the positive ones are empty. Increasing the coupling, i.e. increasing  $\alpha$ , we could expect that the magnitude of the energy is decreasing and the levels corresponding to the bound states will come closer and closer to zero.



With a further increase of the coupling up to a critical value, one possibility for the levels will be just to approach the zero line and never cross. There is, however, also a possibility of crossing. We will see that the first case corresponds to normal spontaneous symmetry breaking. But, if the levels cross, and especially in the most clear case, when they pass the lines  $2m$  and  $-2m$ , respectively, everything will change and we arrive at very different phenomena:

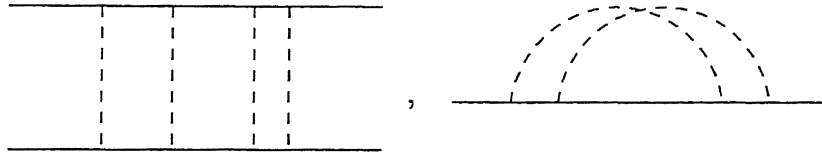


Indeed, now the original vacuum which corresponds to the case when level 2 is empty and level 1 is occupied, is absolutely unstable. We have to fill the new negative energy state and leave the positive energy level empty. But by filling this new state, we get an excitation, a meson-type state with a mass  $\mu$ . For free quarks this would mean that the quark with negative energy decays into a negative energy meson (filling the negative energy levels) and creates a positive energy quark. As a result, the Dirac picture in which all negative energy levels are filled up and all positive energy levels are empty, is destroyed. But if so, a positive energy quark also decays into a positive energy meson and a quark with negative energy. This means that both decays

$$\begin{aligned} q^- &\rightarrow \mu_- + q^+ \\ q^+ &\rightarrow \mu_+ + q^- \end{aligned}$$

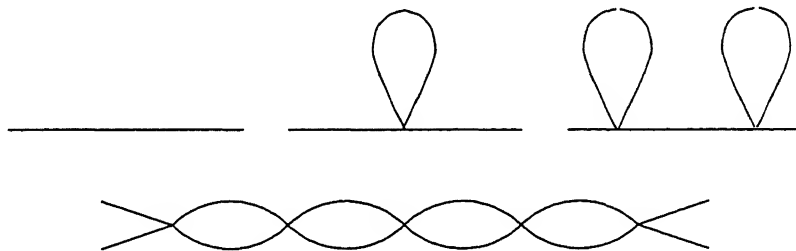
are possible, and both  $q^-$  and  $q^+$  are unstable.

The question is now, how to deal with the bound state problem. Of course, we could just start to calculate the bound states, considering the interactions without corrections to the Green function. However, one has to take into account that the fermion-fermion interaction changes the effective mass of the quarks and this in its turn will change the bound states considerably,

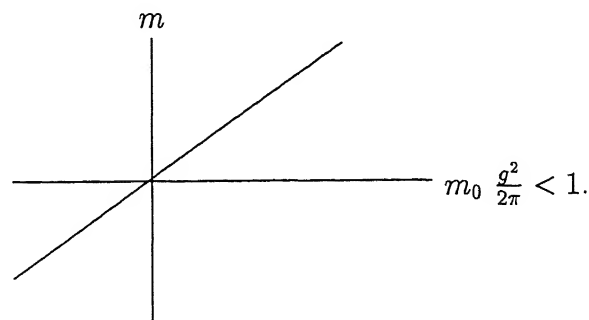


which makes the problem more complicated. We thus will have to consider bound states and the Green functions on equal footing.

Up to now, the only approach which deals with this problem and is self-consistent is the Nambu - Jona-Lasinio model [3]. It considers the fermion Green function corrections due to a four-fermion interaction:



In spite of the strong dependence on the cut-off, the model preserves all symmetries in the Green function and in two-particle interactions. Let us present the result of Nambu and Jona-Lasinio in a way somewhat different from what is given in [3]. We express it as the dependence of the renormalized mass  $m$  on the bare mass  $m_0$ . They found that if the effective coupling (it depends on the definition in their case)  $\frac{g^2}{2\pi}$  is less than unity, the curve will be just the usual one:



If, however, the coupling  $\frac{g^2}{2\pi}$  is larger than unity, the dependence will be like <sup>3</sup>:

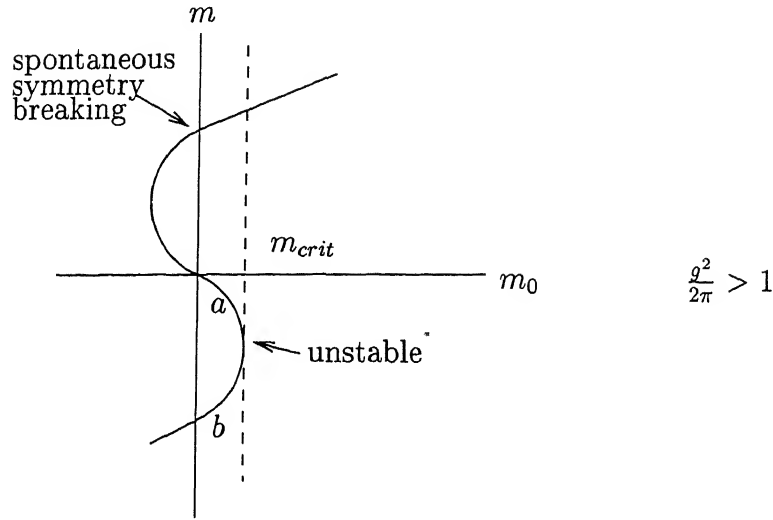


Fig. 2.

According to the interpretation of Nambu and Jona-Lasinio, the upper part of the curve, which at  $m_0 = 0$  reaches a finite point, corresponds to the spontaneous symmetry breaking. But there are three solutions at  $m_0 = 0$  and at sufficiently small  $m_0$  values. What Nambu and Jona-Lasinio claim is that the lower part of the curve is unstable, and there is a real vacuum. I agree with this, if  $m_0 > m_{crit}$ . We can ask: what is the source of instability of this curve? The general argumentation is the following. Talking about spontaneous symmetry breaking,  $m$  is like a magnetic field in a ferromagnetic; we just choose a definite direction. But  $m_0$  is like an external field, and the system is like a compass. If the external field and the induced field are pointing at the same direction, the situation is stable. If they point to opposite directions, the compass will change.

I am, however, not sure that the instability of the almost perturbative solution which contains no condensate at all can be explained in such a way. The explanation may be right for the part  $b$  of the curve in Fig.2 which corresponds to a big spontaneous magnetic moment and the opposite direction of  $m_0$ . It does not work for the part  $a$  close to perturbation theory which has no spontaneous magnetic moment. And, looking more carefully at the curve, we see that the part  $b$  corresponds to pseudoscalar states inside the Dirac sea, while on the piece  $a$  both the pseudoscalar and scalar states are inside, both levels passed. Recognizing this, one can conclude that indeed, the mentioned state is unstable, but for a trivial reason: the corresponding level is inside the Dirac sea and it is not filled up. The problem is, what happens if we fill up this level. It remains an open question, what can be considered as a ground state under these conditions. And it is a problem how to get these results in a more self-consistent way, not depending on the cut-off so strongly.

The Nambu - Jona-Lasinio model can be reproduced in our picture. For this purpose, just as

<sup>3</sup>This result is not always quoted, but it is present in their paper.

a theoretical exercise, let us use  $\alpha$  not going to unity, i.e. draw

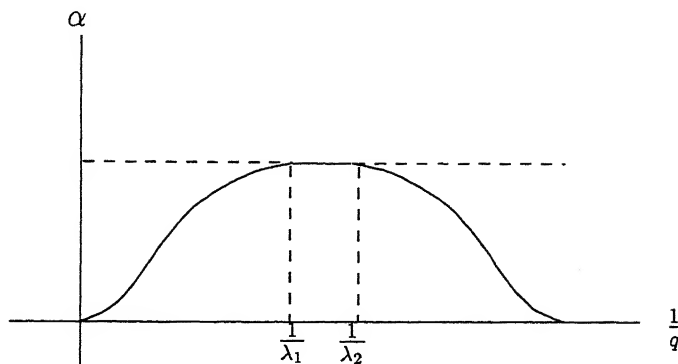
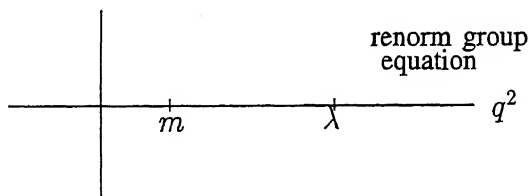


Fig.3.

instead of the curve in Fig.1. In this case there will be a second scale  $\lambda_2$ , and outside this scale we will have just point-like interaction. This reminds of the Nambu - Jona-Lasinio picture, which, apparently, can be reached somehow in our approach. The problem is, how to write constructively the corresponding equation, and whether this can be done at all. Of course, this constructive part can be only approximate. But if we recognize that it can be written, then we will be able to develop a theory in which we put the main ingredient of our discussion as an input into our solution and try to find the real construction. The main difference will appear in the analytic properties of the Green function. The Green function of a fermion for such a case would be quite different in its analytic structure compared to the usual one.

I am afraid I will not have the time to come to this point today, but I would like to explain just the physics.

How to write the equation? What happens in the real case and how to deal with it? Let us start with the Green function. What do we know, what are we supposed to know about this Green function as a function of  $q^2$ ?



Beyond a certain  $\lambda$  in the region where the coupling is small, it is asymptotically free; here the Green function has to satisfy the renorm group equation. But if as a result of the interaction a mass is acquired, this mass would be somewhere at smaller  $q^2$ ; here the equation becomes essentially very complicated and we are not able to extract a reasonable structure.

The idea to write an equation which is correct in both regions, near the threshold and at large  $q^2$ , and to match these two solutions, comes from the following consideration. Suppose that  $\alpha_{crit}/\pi$  is small:

$$\frac{\alpha_{crit}}{\pi} \approx 1 - \sqrt{\frac{2}{3}} = 0.2.$$

Now, however, we may ask: how could new masses, new solutions etc. appear at all at such a small  $\alpha$ . Obviously, this  $\alpha$  has to be multiplied by something large. What happens, for example, at large  $q$ ? We know, that there is always a logarithm of  $q^2/\lambda^2$  and the real parameter becomes

$$\alpha_0 \ln \frac{q^2}{\lambda^2}$$

which is, in spite of the smallness of  $\alpha$ , big enough to change the Green function essentially. But near the threshold there is also a logarithm:

$$\ln \frac{q^2 - m^2}{\tilde{\lambda}^2}, \quad \text{with some scale } \tilde{\lambda} = \lambda \text{ or } m$$

which is always present. In other words, in this region there could be also a quantity which changes seriously in spite of the relative smallness of  $\alpha$ . Hence, we want to write the equation which is correct near the threshold, taking into account correctly the singularity of a supposed mass, and after that compare this with the renorm group equation; we shall see whether it is possible to write an equation which is correct in both regions, and if yes, we will try to solve it. In order to get the singularity correctly, we take the second derivative of  $G^{-1}$  with respect to the momenta.

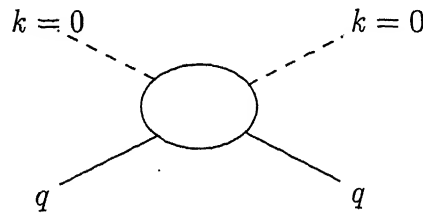
$$G^{-1} = G_0^{-1} + \frac{\frac{1}{(q-q')^2}}{q \quad G(q')} + \frac{\text{diagram with two loops}}{q' \quad q''} \quad (2)$$

The contribution of the first term is trivial, the second derivative of  $q + m$  gives zero. Taking the second derivative of the first graph, it can be easily seen that

$$\partial^2 \frac{1}{(q-q')^2 - i\varepsilon} = -4\pi i \delta^{(4)}(q-q'). \quad (3)$$

This gives for the first diagram  $\gamma_\mu G(q) \gamma_\mu \frac{\alpha}{\pi}$  – just by direct calculation. In other words, we make it local. From this diagram we take the contribution where  $k \equiv (q - q')$  – the momentum of the photon – essentially equals zero.

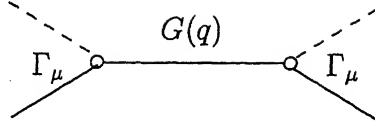
Let us now look at the second diagram. We have here the choice of taking the second derivative at one of the photon lines, or to differentiate once at one line and once at the second line. Having in mind that all the integrations would have a structure which need some logarithmic enhancement, it would mean that the most important regions of integration in this integral would be those where  $k_1 \ll k_2$  and  $k_2 \ll k_1$ . We take the derivative at  $k_1$  and then set it to be zero, but for  $k_2$  the integration will give the same as before. If this integration gives us two logarithms, we kill one and recover it after the integration of our differential equation; but we still have the first one. But if we differentiate once one line and once the other, we will always sit on the region  $k_1 \sim k_2$ , because they have to be of the same order. And, in this case, there is no logarithm at all; after the integration, we will recover one, but one order will be lost. Clearly, a possible approach is to try not to choose different diagrams, but to use the small  $k$  region of integration. Ordering the integration inside the diagram in such a way that one momentum is much smaller than the others, and differentiating this diagram, we will find a relatively simple answer. Indeed, suppose that we have any diagrams with any loops. If we differentiate some lines twice (it can be any line) and neglect all first derivatives, we get an amplitude of the following structure:



This is just the Compton scattering of a zero momentum photon  $k = 0$ , and for this quantity the most singular contribution is obviously

$$\Gamma_\mu(0, q) G(q) \Gamma_\mu(0, q)$$

which corresponds to the diagram



But the vertex  $\Gamma$  is at zero momentum and hence  $\Gamma_\mu(0, q) = \partial_\mu G^{-1}(q)$ . In this approximation we can write a very simple equation:

$$\partial^2 G^{-1}(q) = \frac{\alpha(q)}{\pi} \partial_\mu G^{-1}(q) G(q) \partial_\mu G^{-1}(q) \quad (4)$$

which differs essentially from any Bethe-Salpeter type equation. Indeed, using a Bethe-Salpeter type equation, we do not change the vertex part and end up with rather bad properties. Equation (4) is scale invariant, it is  $\gamma_5$ -invariant, it has many nice symmetry properties and, what is most important, it has a correct behaviour near the threshold.

The gauge is fixed, because we used

$$D_{\mu\nu} = \frac{\delta_{\mu\nu}}{q^2}.$$

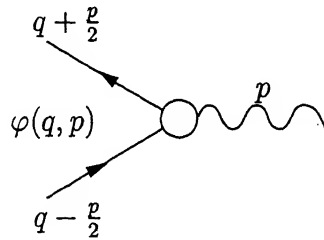
It is an important question, what we would get in different gauges. In Feynman gauge we are very lucky: we find an expression which does not depend explicitly on the expression for the Green function. Using a different gauge, we would find the infrared behaviour of this diagram to be more complicated, and we would not be able to extract universally the region of small  $k$ . We would have integrals over  $q$  which are also possible to use, but with the necessity to think about the behaviour near the threshold.

We, however, have chosen this gauge; we did not destroy the general features and used the current conservation which just corresponds to  $\Gamma_\mu = \partial_\mu G^{-1}$ . Accepting this, we can now ask, what is the relation to the renorm group equation.

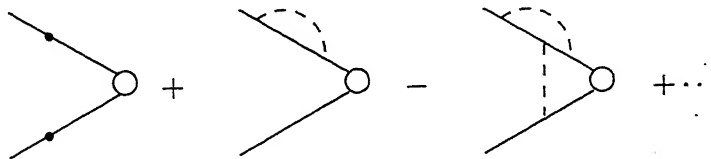
Suppose that we would like to write the renorm group equation in the same spirit. Let us take again the second derivatives. In this case we would be definitely correct, because we know that it is a logarithmic approximation.

In our logarithmic approximation we will do exactly the same with the only difference that  $\alpha$  would be  $\alpha(q^2)$ . In the renorm group equation  $\alpha$  is a function of  $q^2$ . But, of course,  $\alpha$  in general depends on two momenta:  $k^2$  and  $q^2$ . And in the renorm group equation at large momenta, in the ultraviolet region,  $\alpha$  depends on the variable which is the largest. Since we are close to  $k = 0$ , this means that here we will have  $\alpha(q^2)$ , and we will recover the renorm group equation at large  $q^2$ . If we solve this equation with a slowly varying  $\alpha$ , we will be correct in both the threshold region and the ultraviolet region.

We also have to formulate an equation for the bound states under the same assumption. Looking for bound states, we consider scalar and pseudoscalar vertices. This vertex



has to be equal to



Here  $\varphi(q, p)$  depends on  $p$ , the total momentum of a pair, and  $q$ , the quark momenta being given by  $q + p/2$  and  $q - p/2$ . With the same procedure as in obtaining the equation for the Green function we find for the vertex (see [4] for some details):

$$\partial^2 \varphi(q, p) = \frac{\alpha}{\pi} \left[ A_\mu(q) \partial_\mu \varphi(q, p) + \partial_\mu \varphi \tilde{A}_\mu(q) - A_\mu(q) \varphi(q, p) \tilde{A}_\mu(q) \right], \quad (5)$$

where  $A_\mu = \partial_\mu G^{-1} G$ ,  $\tilde{A}_\mu = G \partial_\mu G^{-1}$ . It means that we have two equations in this approximation. We used this approximation just to be constructive and to study what will result if we make this approximation. In principle, solving both equations we will get everything what is necessary: we know  $G$  and  $A_\mu$ , we have a linear equation for bound states, we can see what is the type of the energy etc. The equation for the bound states has very nice features. It is beautiful from the point of view of the Goldstone theorem in the following sense.

Suppose I have some symmetry in my equation, e.g.  $\gamma_5$ -invariance. Since there is no  $\gamma_5$  in equation (5), it is  $\gamma_5$ -invariant. But of course the boundary condition for  $G^{-1}$  at  $q \rightarrow \infty$  is just  $G_0^{-1} = (\hat{q} - m_0)$ , and thus destroys the symmetry. But suppose that  $m_0 = 0$ . In this case there would be symmetry here, which means that the Green function will not be unique, since it can be

$$G^{-1} + \delta G^{-1},$$

where  $G^{-1}$  is some solution and  $\delta G^{-1} \propto \gamma_5 G^{-1}$ . This means that the variation  $\delta G^{-1}$  also is important. What would be the equation for the variation? If we calculate the variation on both sides of equation (4) we obtain

$$\partial_\mu (\delta G^{-1}) = \frac{\alpha}{\pi} (\partial_\mu (\delta G^{-1}) G \partial_\mu G^{-1} + \partial_\mu G^{-1} G \partial_\mu (\delta G^{-1}) - \partial_\mu G^{-1} G (\delta G^{-1}) G \partial_\mu G^{-1}),$$

so we find that  $\varphi = \delta G^{-1}$  fulfils equation (5) at  $p = 0$ . It means that if some symmetry is broken, i.e. if there are multiple solutions of the equation for the Green function, we always will have some solution of the equation for the vertex at  $p = 0$ , which is the Goldstone.

It is clear, that in the present model we can discuss many questions, use a running coupling  $\alpha$  as in Fig. 1 and reproduce the NJL-features without any essentials depending on a cutoff. Before discussing this point further, we will first look for the solution of (4) and discuss the result.

Above, we introduced  $A_\mu = (\partial_\mu G^{-1}) G$ , which is a very useful quantity. Since  $G^{-1} = a \frac{\hat{q}}{q} + b$  is essentially a  $2 \times 2$  matrix,  $A_\mu$  is a  $U(2)$  gauge potential:

$$\partial^2 G^{-1} = \partial_\mu ((\partial_\mu G^{-1}) G G^{-1}) = (\partial_\mu A_\mu) G^{-1} + A_\mu (\partial_\mu G^{-1}) = \frac{\alpha}{\pi} A_\mu \partial_\mu G^{-1} \quad (6)$$

where in the last step we used Eq. (4). Multiplying from the right by  $G$  we thus find

$$\partial_\mu A_\mu + A_\mu A_\mu = \frac{\alpha}{\pi} A_\mu A_\mu. \quad (7)$$

This means that

$$\partial_\mu A_\mu = -\beta A_\mu A_\mu;$$

and  $A_\mu$  is a pure  $U(2)$ -gauge potential with a condition  $\beta = 1 - \frac{\alpha}{\pi}$ . Of course, this is just a useful trick. Important is to write down the real equation for the Green function. The most natural thing is to express  $G^{-1}$  in the form

$$G^{-1} = \rho e^{\frac{\alpha}{2} \hat{n}},$$

where  $\hat{n}$  is a  $2 \times 2$ -matrix

$$\hat{n} = \frac{\hat{q}}{q}.$$

It is just easier to use this form for our purpose: we can find an equation for  $\rho$  and an equation for  $\varphi$ . Both are functions of  $q^2$ :  $\rho(q^2)$ ,  $\varphi(q^2)$ . There is, however, no scale in the equation; it contains only a derivative of  $q$ . If we introduce

$$\xi = \ln q,$$

we will find an equation in which  $\xi$  can be considered as "time", and which is an oscillator equation. In fact there are two oscillators, one for  $\rho$ , the other for  $\varphi$ , and they will satisfy non-linear equations. For  $\varphi$  we find

$$\ddot{\varphi} + 2 \left( 1 + \beta \frac{\dot{\rho}}{\rho} \right) \dot{\varphi} - 3 \sinh \varphi = 0. \quad (8)$$

This is just an oscillator with damping; a similar equation can be written for  $\rho$ . Important is that that there has to be "energy" conservation in this equation. Indeed, we said that  $\xi$  plays the role of time; it, however, did not enter the equation explicitly. Thus there has to be a conservation law which, as it is easy to show, leads to

$$\left( 1 + \beta \frac{\dot{\rho}}{\rho} \right)^2 = 1 + \beta^2 \left( \frac{\dot{\varphi}^2}{4} - 3 \sinh^2 \frac{\varphi}{2} \right). \quad (9)$$

We thus can eliminate  $\rho$  altogether, and find the equation for  $\varphi$

$$\ddot{\varphi} + 2 \sqrt{1 - \beta^2 \left( 3 \sinh^2 \frac{\varphi}{2} - \frac{\dot{\varphi}^2}{4} \right)} \dot{\varphi} - 3 \sinh \varphi = 0,$$

which is an oscillator with damping. Having this in mind is sufficient to understand the structure of the solution. Indeed, what is this  $\varphi$ ? We have

$$G^{-1} = \rho \cosh \frac{\varphi}{2} + \frac{\hat{q}}{q} \rho \sinh \frac{\varphi}{2}. \quad (10)$$

The perturbative solution is  $\varphi$  close to  $i\pi$ . In this case the first term is zero, the other is proportional to  $\hat{q}/q$  - this corresponds to the massless solution. Since  $m_0$  is small, we have to have solutions like this at  $q \rightarrow \infty$ .

Now we have to find the solution everywhere. Let us first investigate the equation without damping; we get

$$\ddot{\varphi} - 3 \sinh \varphi = 0.$$

If we go to the Euclidean space,  $\varphi = i\psi$ , the potential becomes a periodical potential:

$$\ddot{\psi} - 3 \sin \psi = 0.$$

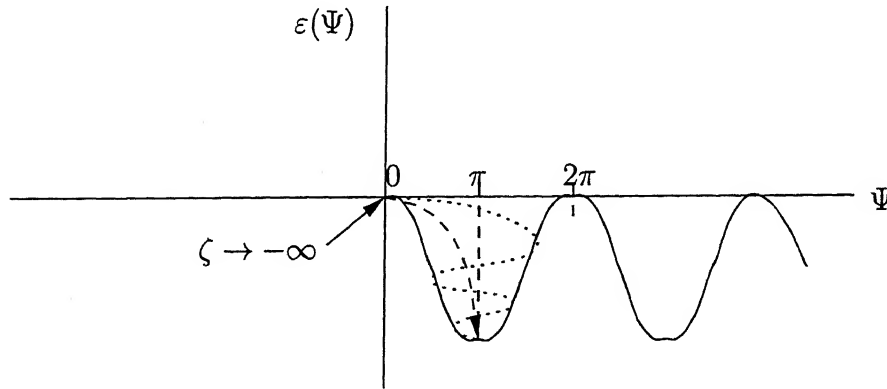


Fig. 4.

We have to look for a possible solution for this structure with damping. What does this mean? For the oscillator with damping any solution at  $\xi \rightarrow \infty$  has to be in a minimum, because the energy is decreasing. But if  $\xi$  is going to  $-\infty$ , the energy is growing. What could be in this case a normal, reasonable solution? It is almost clear that the only possibility is to put at  $\xi \rightarrow -\infty$  the "particles" in this oscillator at the maximum, and start to move them slowly; eventually, they will appear inside the well.



There is a most important question, namely: what is the critical coupling in this case? What do we know about an oscillator with damping? If the damping is large enough, all the trajectories will go monotonically to the minimum. If the damping is sufficiently small, the solution will start to oscillate. In order to see when this transition happens, we have to look for the equation just near the minimum  $\psi = \pi$ . With  $\phi \equiv \pi - \psi$  we have for small  $\phi$

$$\ddot{\phi} + 2\sqrt{1 + 3\beta^2}\dot{\phi} + 3\phi = 0,$$

with fundamental solutions

$$\phi_{1,2} = e^{\nu_{1,2}\xi} \text{ where } \nu_{1,2} = -\sqrt{1 + 3\beta^2} \pm \sqrt{3\beta^2 - 2}.$$

So we have monotonic behaviour for  $3\beta^2 - 2 > 0$ . On the other hand if  $\beta^2 < \frac{2}{3}$ , i.e.

$$\frac{\alpha_{crit}}{\pi} = 1 - \sqrt{\frac{2}{3}} < \frac{\alpha}{\pi} < 1 + \sqrt{\frac{2}{3}},$$

we will have oscillations before reaching the minimum. The critical angle  $\psi_c$ , which separates the regions where the solution is monotonic and where it oscillates can be shown (see e.g. (4)) to be given by

$$\sin^2 \frac{\psi}{2} = \left( \frac{2}{3} - \beta^2 \right) \sqrt{\frac{1 + 3\beta^2}{1 - \beta^2}} \frac{1}{1 + \sqrt{(1 + 3\beta^2)(1 - \beta^2)}}.$$

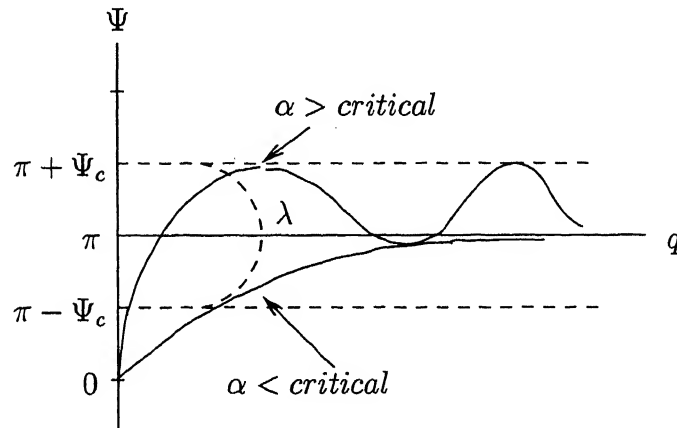


Fig. 5.

Up to now we have considered a constant coupling  $\alpha$ . We know that for  $q > \lambda$  the Green function is determined by perturbation theory, which has to match the solutions in the region of smaller  $q$ . If  $\beta^2 > \frac{2}{3}$  for all  $q^2$ , the solution which goes as  $\psi \approx \frac{q}{m_c}$  for  $q \rightarrow 0$  matches the solution  $\frac{i}{2}(\psi - \pi) \approx \frac{m_0}{q} + \frac{\nu_1^2}{q^3}$  for  $q \rightarrow \infty$  monotonically. This determines  $m_0$  as a function of  $m_c$  in a unique way. Let  $\lambda$  be the value of  $q$  where  $\beta^2(\lambda^2) = \frac{2}{3}$ . If, however,  $\beta^2 < \frac{2}{3}$  below  $q = \lambda$ , the solutions can

oscillate and  $m_0(m_{c_i}) = 0$  for some  $m_{c_i}$  as indicated in Fig.6.

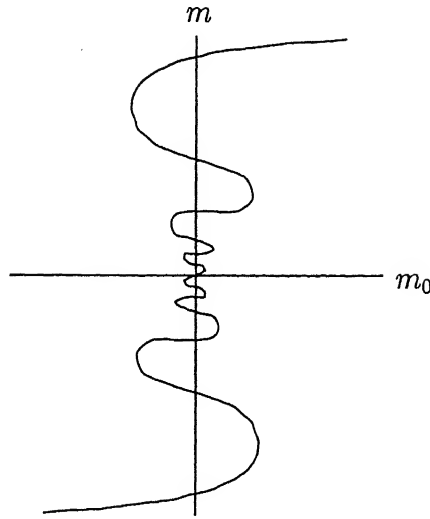


Fig.6

This then is a solution corresponding to broken chiral symmetry.

## References

- 1 V. N. Gribov, Orsay lectures on confinement (I), preprint LPTHE Orsay 92-60 (1993); hep-ph/9403218
- 2 V. N. Gribov, Orsay lectures on confinement (II), preprint LPTHE Orsay 94-20 (1994); hep-ph/9407269
- 3 Y. Nambu, G. Jona-Lasinio, Phys. Rev. 122 (1965), 345
- 4 V. N. Gribov, Lund preprint LU 91-7 (1991)
- 5 V. N. Gribov, QCD at large and short distances, Bonn preprint TK 97-08 (1997), hep-ph/9807224.
- 6 V. N. Gribov, The theory of quark confinement, Bonn preprint TK 98-09 (1998), hep-ph/9902279

# 7. An Essay on Color Confinement

Kazuhiko Nishijima <sup>(a)</sup> \* and Masud Chaichian <sup>(b)</sup> †

(a) Nishina Memorial Foundation

2-28-45 Honkomagome, Bunkyo-ku, Tokyo 113-8941, Japan

(b) Helsinki Institute of Physics

P.O. Box 9, FIN-00014 Helsinki, Finland

## Abstract

Color confinement is a consequence of an unbroken non-Abelian gauge symmetry and the resulting asymptotic freedom inherent in quantum chromodynamics. A qualitative sketch of its proof is presented.

## 1 Introduction

There has been an accumulation of evidence in favor of the quark model of hadrons [1] and we can no longer think of any other substitute for it. Yet, no isolated quarks have been observed to date, and we are inclined to think that observation of isolated quarks is, in principle, impossible. This is the hypothesis of quark confinement, and it has been further extended to that of color confinement that implies not only the unobservability of quarks but also of all the isolated colored particles such as quarks and gluons. Then a natural question is raised of whether or not we can account for this hypothesis within the framework of the conventional quantum chromodynamics (QCD) dealing with the gauge interactions of quarks and gluons. The answer to this question is affirmative and the detailed mathematical proof of color confinement has been published elsewhere [2-6]. In this article, therefore, we shall follow the flow of ideas underlying the proof in a qualitative manner.

The problem of color confinement may be decomposed into two steps. The first step consists of finding a consensus of interpretations of color confinement. Unless it is properly settled we do not know what we have to prove in the second step. Because of the importance of this subject many authors have proposed various interpretations. A typical example is Wilson's area law for the loop correlation function in the lattice gauge theory [7]. When it is obeyed the interaction between a quark and an antiquark is given by a confining linear potential. Another example is given by coherent superposition of magnetic monopoles in the vacuum state [8-13]. This is dual to the superconducting vacuum based on coherent superposition of charged objects such as the Cooper pairs. Corresponding to the superconductor of the second kind a pair of magnetic monopoles can be connected by a quantized magnetic flux forming a hadronic string whose energy is proportional to the distance between them. Then the situation is similar to the preceding example.

In these examples one introduces a topological structure through monopoles, strings and instantons into the configuration space. In the present paper, however, we shall consider a different topological structure in the state vector space. For this purpose we look for a known example of confinement within the framework of known field theories, and we find a prototype example in quantum electrodynamics (QED) [2]. When the electromagnetic field is quantized in a covariant gauge, say, in the Fermi gauge, three kinds of photons emerge, namely, transverse, longitudinal and scalar photons, but only the transverse photons are subject to observation leaving the other two unobservable. We recognize that this is indeed a typical example of confinement, and we may

---

\*E-mail: nisijima@argus01.phys.su-tokyo.ac.jp

†Email: chaichia@pcu.helsinki.fi

be able to find some clues to color confinement by studying closely the mechanism of confinement of longitudinal and scalar photons in QED. For this reason we analyze its mechanism in Sec. 2 so that we can generalize it and apply it to QCD.

One of the profound features of gauge theories is the Becchi-Rouet-Stora (BRS) invariance [14] and its introduction is vital to the interpretation of confinement. Therefore, we shall describe some of the basic properties of this invariance in Sec. 3.

The strong interactions described by QCD possess a novel feature called asymptotic freedom [15,16], and in Sec. 4 we shall discuss how this aspect of strong interactions drew our attention and how the non-Abelian gauge theory entered the game. Finally in Sec. 5 we shall combine BRS invariance with asymptotic freedom to prove color confinement.

## 2 Quantum Electrodynamics and Indefinite Metric

When the electromagnetic field is quantized in a covariant gauge, say, in the Fermi gauge, we find transverse, longitudinal and scalar photons, but the latter two are never observed. We may interpret it as an example of confinement, and we have at least three alternative ways of explaining it. First, we can refer to the representations of the Poincaré group for massless particles [17,18]. Then, massless particles are known to have only two directions of polarization no matter what their spin is. Thus photons are always transversely polarized and the same would be true with gluons if they could be observed. The second method is to employ the Coulomb gauge by keeping only the transverse photons from the start. The remnants of unobservable photons manifest themselves in the form of the Coulomb potential. This method is applicable, however, only to the linear Abelian gauge theories such as QED. The third and the most useful method is the introduction of a subsidiary condition such as the Lorentz condition.

Quantization of the electromagnetic field in a covariant gauge forces us to introduce indefinite metric [19] which is inherited from the Minkowski metric. Thus the whole state vector space in QED can no longer possess the positive-definite metric, and for the physical interpretation of the theory we have to eliminate indefinite metric by imposing the Lorentz condition on the state vectors to select observable or physical states. In order to execute this program let us quantize the free electromagnetic field in the Fermi gauge and for a given momentum we have four directions of polarization, namely, two transverse, one longitudinal and one scalar. Thus we have four kinds of photons specified by the directions of polarization. The canonical quantization then implies that the scalar photons are represented by negative norm states. This is a consequence of the manifest covariance of the quantization of the vector field in the Minkowski space.

The emergence of indefinite metric indicates that observable states occupy only a portion of the whole state vector space called the physical subspace. In order to define such a subspace we introduce a subsidiary condition known as the Lorentz condition. Let us consider the four-divergence of the vector field, then it represents a free massless field even in the presence of the interactions. We decompose it into a sum of positive- and negative- frequency parts corresponding to destruction and creation operators, respectively. We find that the photons involved in this operator are special combinations of the longitudinal and scalar photons in the amplitude. We shall call them a-photons, then an a-photon state has zero norm. We can introduce an alternative combination of longitudinal and scalar photons called b-photons in such a way that a b-photon state also has zero norm. Thus for a given momentum we have two transverse (t-) photons, an a-photon and a b-photon. Although both an a-photon state and a b-photon state have zero norm, their inner product is non-vanishing so that they are metric partners.

A physical state is defined as such a state that is annihilated by applying the positive frequency part of the four-divergence of the vector field. This is the Lorentz condition. We can easily verify that the S matrix in QED transforms a physical state into another physical state since it commutes with the four-divergence. This is one of the general features of the subsidiary condition. Also we can easily verify that the b-photons are excluded from the physical subspace. Therefore, we have only t-photons and a-photons in the physical states. Then we can show that the inner product of a physical state involving at least one a-photon with another physical state vanishes identically. In

other words, a-photons give no contributions to observable quantities, and both a- and b-photons escape detection. This is the confinement mechanism of the longitudinal and scalar photons. In QED only the transverse photons remain observable. In QCD, however, not only longitudinal and scalar gluons but also transverse gluons are unobservable. Thus, there are some essential differences in the nature of confinement between QED and QCD. In the former case confinement is kinematical in the sense that it could be understood without recourse to dynamics of the system, whereas in the latter case it is dynamical in nature as the proof depends sensitively on the dynamical properties of the system.

### 3 Quantum Chromodynamics and BRS Invariance

As we shall see in the next section strong interactions of quarks are mediated by a non-Abelian gauge field corresponding to the  $SU(3)$  color symmetry. Thus we shall discuss one of the most characteristic features of gauge theories known as the BRS invariance in this section [14].

Classical electrodynamics is gauge-invariant. Field strengths expressed in terms of the vector field are invariant under the local or space-time-dependent gauge transformations of the latter. Given a source term, therefore, the solution of the equation for the vector field is not uniquely given, and this non-uniqueness is an obstacle to quantization. In order to overcome this difficulty we add to the gauge-invariant Lagrangian a term violating the local gauge invariance. This extra term is called the gauge-fixing term and was first introduced by Fermi. Later it has been generalized so as to include an arbitrary parameter called the gauge parameter. In the original form introduced by Fermi this parameter is equal to unity.

After quantization we find that we have to introduce indefinite metric into the state vector space and that the divergence of the vector field commutes with the  $S$  matrix. Because of the inclusion of the gauge-fixing term the field equation deviates from the classical Maxwell equation by a term proportional to the four-divergence of the vector field. It so happens that a matrix element of this four-divergence between two physical states vanishes identically because of the Lorentz condition, and the classical Maxwell equation is recovered in the physical subspace. In this way we find, despite the introduction of the gauge-fixing term, that expectation values of gauge-invariant quantities and the  $S$  matrix elements in the physical subspace are independent of the choices of the gauge parameter because of the congeniality between the gauge-fixing term and the subsidiary condition. In what follows we shall extend this approach to QCD.

There are many essential differences between QED and QCD, however. The former is an Abelian gauge theory described by a linear field equation, whereas the latter is a non-Abelian gauge theory described by a non-linear field equation. In both cases the gauge-invariant part of the Lagrangian is given by the square of the field strength. So, let us introduce the gauge-fixing term in QCD assuming the same structure as in QED. Then we recognize that it does not work because observable quantities depend explicitly on the gauge parameter. Another difficulty arises from the fact that the four-divergence of the gauge field is no longer a free field, and this prevents us from defining its positive frequency part. In other words, the Lorentz condition cannot be employed to define physical states in QCD. Thus we are obliged to find a device to overcome these difficulties and to this end we shall introduce the Faddeev-Popov ghost fields.

In order to eliminate the gauge-dependence of physically relevant quantities Faddeev and Popov have proposed a procedure of averaging the path integral over the manifold of gauge transformations. We skip the mathematical detail here and refer to the original paper [20], but we should mention that this procedure resulted in a new additional term in the Lagrangian called the Faddeev-Popov (FP) ghost term. This term involves a pair of Hermitian scalar fields, but they are anticommuting and consequently violate Pauli's theorem on the connection between spin and statistics. For this reason they are called ghost fields. Pauli's theorem is based on three postulates, (1) Lorentz invariance, (2) local commutativity or microscopic causality and (3) positive-definite metric for state vectors, and the FP ghost fields violate the last one obliging us to introduce indefinite metric into the theory.

Thus we face again the problem of eliminating indefinite metric from the theory with the help

of an appropriate subsidiary condition to select physical states out of the whole state vector space. When physical states are so defined as those that are annihilated by applying a certain operator, that operator should commute with the  $S$  matrix as does the four-divergence of the vector field in QED. In order to find such an operator a novel symmetry discovered by Becchi, Rouet and Stora is extremely useful. Although this symmetry was originally utilized in renormalizing QCD, it plays an essential role in the proof of color confinement in QCD.

In a classical gauge theory a local gauge transformation is specified by a function of the space-time coordinates called the gauge function and the classical theory is invariant under such a transformation. This local gauge invariance is lost when the gauge-fixing and FP ghost terms are introduced. Besides, local gauge transformations are defined only for the color gauge field and the quark fields, but they are not even defined for FP ghost fields. The BRS transformations for the color gauge field and the quark fields are given by replacing the gauge function by one of the FP ghost fields in infinitesimal gauge transformations. Since we have a pair of ghost fields we introduce, correspondingly, a pair of BRS transformations. Then a question is raised of how to define BRS transformations of the ghost fields since their gauge transformations are not defined. Fortunately, this problem has a simple but beautiful solution. Their BRS transformations are introduced by demanding the invariance of the total Lagrangian under them.

The total Lagrangian including the gauge-fixing and FP ghost terms is no longer invariant under local gauge transformations, but it is invariant under the global BRS transformations. Noether's theorem then tells us that there must be a pair of conserved quantities corresponding to a pair of BRS symmetries. They are Hermitian and called the BRS charges. As mentioned before there are two kinds of Hermitian FP ghost fields and correspondingly a BRS charge must involve one of the ghost fields. In what follows we keep only one of these two charges for simplicity. The BRS charge that we keep is anticommuting just as the FP ghost field, and consequently the square of the BRS charge vanishes and it is called nilpotent. The Hermiticity and nilpotency of the BRS charge would imply indefinite metric since otherwise it would be a null operator [21,22]. The nilpotency is important and allows us to introduce the concept of cohomology in the theory. After a long detour we are going to introduce an appropriate subsidiary condition. Physical states are defined as those states that are annihilated by applying the BRS charge [23].

The FP ghost fields do not appear in the conventional QED but we can also introduce them although they are non-interacting fields. Then we can combine the Lorentz condition with the additional condition implying the absence of FP ghosts to define the physical states. When these conditions are satisfied, we can prove that physical states so defined are annihilated by the BRS charge in QED.

The BRS charge is the generator of the BRS transformation and the BRS transform of an operator is given by the commutator or anticommutator of that operator with the BRS charge, and this transformation is also nilpotent. An operator which is the BRS transform of another operator is called an exact operator, then it is clear that the matrix element of an exact operator between a pair of physical states vanishes.

The equation for the non-Abelian gauge field deviates from the classical Maxwell equation and in fact the divergence of the field strength plus the color current does not vanish but is equal to a certain exact operator, which will be referred to as an exact current hereafter. Therefore, the classical Maxwell equation is recovered when we take the matrix element of the field equation between a pair of physical states. Furthermore, the BRS charge commutes with the  $S$  matrix. Thus the scenario in QED is reproduced almost exactly.

When single quark states and single gluon states are unphysical these particles are unobservable and consequently confined. Thus the problem of color confinement reduces to that of proving that they are unphysical states. We shall evaluate the expectation value of the exact current in a single quark state or a single gluon state. If they should belong to physical states the expectation values in these states would vanish identically, so that non-vanishing of the expectation values would be a direct indication that these particles are unphysical and confined.

The four-divergence of the exact current vanishes, and we can give a set of Ward-Takahashi identities for Green's functions involving the exact current [2-4]. By making use of the above set of Ward-Takahashi identities we can prove that the expectation value of the exact current in a

single colored particle state survives when the exact current as applied to the vacuum state does not generate a massless spin zero particle. Therefore, the absence of such a massless particle is a sufficient condition for color confinement [2-4]. In order to check its absence we introduce the vacuum expectation value of the time-ordered product of the gauge field and the exact current and evaluate the residue  $C$  of the massless spin zero pole of the Fourier transform of this two-point function. The four-divergence of this two-point function is proportional to this constant  $C$  except for a trivial kinematical factor, and the divergence can be cast in the form of an equal-time commutator.

By checking this equal-time commutator closely we find that  $C$  is the sum of a constant  $a$  and the Goto-Imamura-Schwinger (GIS) term. The constant  $a$  is equal to the inverse of the renormalization constant of the color gauge field. These constants  $C$  and  $a$  satisfy distinct renormalization group (RG) equations and boundary conditions. We shall not enter this subject here since the mathematical detail has been given elsewhere [2-5], but we infer the fact that vanishing of  $a$  automatically leads to vanishing of  $C$  and color confinement is realized. Indeed, it has been known for some time that gluons are confined when  $a$  vanishes [24,25], but now with the help of the BRS invariance we could conclude that not only gluons but also all the colored particles are simultaneously confined. We shall come back to this subject again in Sec. 5.

## 4 Asymptotic Freedom

In this section we shall review briefly how and why our attention was drawn to the non-Abelian gauge theory in describing strong interactions. In particle physics strongly interacting particles such as nucleons and pions are called hadrons. Hadrons are composite particles of quarks and antiquarks, however, and we have to study the origin of the strong interactions of quarks.

We already know that strong interactions are mediated by the color gauge field and the quanta of this field are called the gluons since they glue up quarks together to form hadrons. Dynamics of quarks and gluons is called QCD as mentioned before. In the sixties experiments on the deep inelastic scattering of electrons on protons had been carried out. The differential cross-section had been measured by specifying the energy and direction of electrons without observing the hadrons in the final states. Then, apart from kinematical factors this differential cross-section can be expressed as a linear combination of two structure functions. They are functions of the square of the momentum transfer and the energy loss of the electron in the laboratory system. When these two variables increase indefinitely the two structure functions tend to be functions of the ratio of these two variables except for trivial kinematical factors. This characteristic behavior of structure functions is called the Bjorken scaling [26], and it is considered to be an empirical manifestation of the properties of strong interactions. What do we learn from this? In 1969 Feynman proposed the parton model and assumed that a nucleon consists of point-like partons moving almost freely inside the nucleon [27]. In order to keep the partons inside the nucleon, however, the four-momentum of a parton must be equal to a fraction  $x$  of the total four-momentum of the nucleon. The partons may be identified with the quarks and since  $x$  is identified with the ratio of the two kinematical variables referred to in the above the distribution of the fraction  $x$  has been shown to be related to the structure functions.

From the success of the parton model in reproducing the Bjorken scaling we may infer that quarks inside the hadrons are almost free and that the interactions of quarks turn out to be weaker at shorter distances. This is a distinctive feature of strong interactions and we may express it in the momentum space as follows: The probability of a process involving large momentum transfer in strong interactions is small.

We look for a model satisfying this condition and find that only non-Abelian gauge interactions meet this requirement with the help of RG [15,16].

The concept of RG was first introduced by Stueckelberg and Petermann in 1953 [28], and it was further advanced by Gell-Mann and Low in QED in 1954 [29]. Let us consider a dielectric medium and put a positive test charge inside, then the medium is polarized, namely, negative charges are attracted and positive ones are repelled by this test charge. As a consequence it induces a new

charge distribution in the medium. The total charge inside a sphere of radius  $r$  around the test charge is a function of  $r$  and we call it the running charge. The vacuum is an example of the dielectric media because of its ability of being polarized – the vacuum polarization. In this case the test charge is called the bare charge and the total charge inside a sphere of a sufficiently large radius is called the renormalized charge. The running charge is a function of the radius  $r$ , but it can also be regarded as a function of momentum transfer through the Fourier transformation. The bare charge then corresponds to the limiting value of the running charge for infinite momentum transfer.

Gell-Mann and Low have proved on the basis of the RG method that given a finite renormalized charge the bare charge is equal to a certain finite constant independent of the value of the renormalized one or it is divergent [29]. The Bjorken scaling phrased in terms of RG implies that the bare coupling constant must be equal to zero. We shall refer to this property as asymptotic freedom (AF), and the non-Abelian gauge theory is the only known example in which AF is realized as clarified by Gross and Wilczek and by Politzer [15,16]. The origin of AF may be traced back to the fact that the vector field introduces indefinite metric needed to realize AF and that the non-Abelian gauge theory is the only example involving non-linear interactions of the vector field.

Thus starting from the empirical Bjorken scaling we have finally reached the non-Abelian gauge theory of strong interactions, namely, QCD.

## 5 Color Confinement

Now we are ready to present the proof of color confinement, at least verbally, by combining arguments given in preceding sections.

In QED the square of the ratio of the renormalized charge to the bare one is equal to the renormalization constant of the electromagnetic field. It is equal to the inverse of the dielectric constant of the vacuum relative to the empty geometrical space. Usually the dielectric constant of a dielectric medium is defined relative to the vacuum, but here we define it relative to the empty geometrical space or the void.

This dielectric constant of the vacuum is larger than unity as a consequence of the positive-definite metric of the physical subspace, or more intuitively, it is a consequence of the screening effect due to the vacuum polarization. Then, let us consider a fictitious case in which the dielectric constant of the vacuum is smaller than unity. In this case we have antiscreening instead of screening when a test charge is placed in this fictitious vacuum, and such a vacuum is realized when a pair of virtual charged particles of indefinite metric should contribute to the vacuum polarization. In this case the running charge would be an increasing function of the radius  $r$  at least for small values of  $r$ . Next we shall consider an extreme case of the vanishing dielectric constant, then a small test charge would attract an unlimited amount of like charges around it thereby bringing the system into a catastrophic state of infinite charge. Nature would take safety measures to prevent such a state from emerging, and a possible resolution is to bring another test particle of the opposite charge. The total charge of the whole system is equal to zero and charge confinement would be realized. In QED, however the dielectric constant of the vacuum or the inverse of the renormalization constant is larger than unity, and the above scenario reduces to a mere fiction.

The situation in QCD is completely different since it allows introduction of indefinite metric in the vacuum polarization and AF is one of its manifestations. In QCD what corresponds to the dielectric constants of the vacuum in QED is the inverse of the renormalization constant of the color gauge field denoted by  $a$  in Sec. 3. If  $a$  should vanish we would encounter a scenario similar to the one mentioned above and a test color charge would induce an intolerable catastrophic state. In Sec. 3 we have shown that such a state is excluded by means of the subsidiary condition that selects physical states. Therefore, what can be realized are states of zero color charge and this is precisely color confinement. Unlike electric charge, color charge is not a simple additive quantum number but a member of a Lie algebra  $su(3)$ , so that physically realizable states should belong to the one-dimensional representation of this algebra. Thus the entire problem of color confinement



reduces to the proof that the constant  $a$  vanishes.

Before presenting its proof we have to introduce the concept of the equivalence class of gauges [2,4,5]. When the difference between two Lagrangian densities is an exact operator we say that these two Lagrangian densities belong to the same equivalence class of gauges. For instance, two Lagrangian densities corresponding to two distinct values of the gauge parameter belong to the same equivalence class. In QCD hadrons are represented by BRS invariant composite operators [30-32], and the S matrix elements for hadron reactions are obtained by applying the reduction formula of Lehmann, Symanzik and Zimmermann [33] to Green's functions defined as the vacuum expectation values of the time-ordered products of the BRS invariant composite operators. Then we can readily prove that the S matrix elements for hadron reactions are the same within the same equivalence class of gauges [2,4,5]. Color confinement signifies that the unitarity condition for the S matrix in the hadronic sector is saturated by hadronic intermediate states. That means that quarks and gluons have no place to show up in the unitarity condition just as longitudinal and scalar photons never appeared in the S matrix elements in QED. Therefore, we may take it for granted that the concept of color confinement is gauge-independent within the same equivalence class.

Then we come back to the evaluation of the constant  $a$ . First, it should be stressed that  $a$  can be evaluated exactly as a function of the gauge coupling constant and the gauge parameter thanks to AF [2,5]. These two parameters define a two-dimensional parameter space, which is then decomposed into three domains according to the value of  $a$ , namely, zero, infinity and finite. It should be stressed here that the existence of these three domains can be proved without recourse to perturbation theory. Of these three domains color confinement is manifestly realized in the first one, and also in the other two confinement should prevail because of the gauge-independence of the concept of color confinement. Evaluation of  $a$  by means of RG based on AF is a very interesting mathematical problem, but we shall refer to the original paper for the technical detail[5].

Finally, it should be stressed that confinement as has been discussed in this paper is realized only when we have an unbroken non-Abelian gauge symmetry [2]. When a certain gauge symmetry is spontaneously broken the exact current generates a massless spin zero particle as Nambu-Goldstone boson and our proof of confinement breaks down. For instance, the electroweak interactions are formulated on the gauge group  $SU(2) \times U(1)$ , but spontaneous symmetry breaking reduces the gauge symmetry to the Abelian  $U(1)$  corresponding to the electromagnetic gauge symmetry. Thus the electroweak interactions do not possess any unbroken non-Abelian gauge symmetry and are not capable of confining any particle.

To conclude, we have presented the flow of ideas towards intuitive understanding of the mechanism of color confinement without recourse to mathematical detail, but interested readers are encouraged to refer to the original articles.

The authors are grateful to Professor A. N. Mitra for kindly inviting us to contribute this article to the INSA book.

## References

- [1] M. Gell-Mann, Phys. Letters **8**, 214 (1964).
- [2] The most extensive review article on the proof of color confinement in the context of present approach is found in the article, K. Nishijima, Czech. J. Phys. **46**, 1 (1996).
- [3] K. Nishijima, Int. J. Mod. Phys. **A9**, 3799 (1994).
- [4] K. Nishijima, Int. J. Mod. Phys. **A10**, 3155 (1995).
- [5] K. Nishijima and N. Takase, Int. J. Mod. Phys. **A11**, 2281 (1996).
- [6] K. Nishijima, Int. J. Mod. Phys. **B12**, 1355 (1998).
- [7] K. G. Wilson, Phys. Rev. **D14**, 2455 (1974).

- [8] H. B. Nielsen and P. Olesen, Nucl. Phys. **B6** , 45 (1973).
- [9] Y.Nambu, Phys. Rev. **D10** , 4262 (1974).
- [10] S. Mandelstam, Phys. Reports. **C23** , 245 (1976).
- [11] G. 't Hooft, Nucl. Phys. **B79** , 276 (1974).
- [12] N. Seiberg and E. Witten, Nucl. Phys. **B426** , 19 (1994).
- [13] N. Seiberg and E. Witten, Nucl. Phys. **B431** , 484 (1994).
- [14] C. Becchi, A. Rouet and R. Stora, Ann. Phys. **98** , 287 (1976).
- [15] D. J. Gross and F. Wilczek, Phys. Rev. Lett. **30** , 1343 (1973).
- [16] H. D. Politzer, Phys. Rev. Lett. **30** , 1346 (1973).
- [17] E. P. Wigner, Nuovo Cimento **3** , 517 (1956).
- [18] E. P. Wigner, Rev. Mod. Phys. **29** , 255 (1957).
- [19] N. Nakanishi, Prog. Theor. Phys. Suppl. No. 51, 1 (1972).
- [20] L. D. Faddeev and V. N. Popov, Phys. Lett. **25B** , 29 (1967).
- [21] K. Nishijima, Nucl. Phys. **B238** , 601 (1984).
- [22] K. Nishijima, Prog. Theor. Phys. **80**, 897 (1988).
- [23] T. Kugo and I. Ojima, Prog. Theor. Phys. Suppl No. 66, 1 (1979).
- [24] K. Nishijima, Prog. Theor. Phys. **75**, 1221 (1986).
- [25] R. Oehme, Phys. Lett. **195B**, 60 (1987).
- [26] J. D. Bjorken, Phys. Rev. **179**, 1547 (1969).
- [27] R. P. Feynman, Phys. Rev. Lett. **23**, 1415 (1969).
- [28] E. C. G. Stueckelberg and A. Petermann, Helv. Phys. Acta **26**, 499 (1953).
- [29] M. Gell-Mann and F. E. Low, Phys. Rev. **95**, 1300 (1954).
- [30] R. Haag, Phys. Rev. **112**, 669 (1958).
- [31] K. Nishijima, Phys. Rev. **111**, 995 (1958).
- [32] W. Zimmermann, Nuovo Cimento **10**, 597 (1958).
- [33] H. Lehmann, K. Symanzik and W. Zimmermann, Nuovo Cimento **1**, 205 (1955).



# Part B : Topological Aspects Of QFT

8. Topological Quantum Field Theories - A Meeting Ground For Physicists And Mathematicians by R.Kaul
9. Quantum Field Theory And The Jones Polynomial by Edward Witten (Commun.Math Phys. **121**,351-399; 1989) (reproduced under permission from Springer-Verlag)
10. Chiral Anomalies In Field Theories by H.Banerjee
11. Coherent States In Field Theory by Wei-Min Zhang
12. Pancharatnam, Bargmann And Berry Phases - A Retrospective by N.Mukunda
13. The Skyrme Model For Baryons by J.Schechter and H.Weigel



# 8. Topological QFT – A Meeting Ground for Physicists and Mathematicians

Romesh K. Kaul \*

The Institute of Mathematical Sciences,  
Taramani, Chennai 600 113, India

## Abstract

Topological quantum field theories can be used as a powerful tool to probe geometry and topology in low dimensions. Chern-Simons theories, which are examples of such field theories, provide a field theoretic framework for the study of knots and links in three dimensions. These are rare examples of quantum field theories which can be exactly (non-perturbatively) and explicitly solved. Abelian Chern-Simons theory provides a field theoretic interpretation of the linking and self-linking numbers of a link. In non-Abelian theories, vacuum expectation values of Wilson link operators yield a class of polynomial link invariants; the simplest of them is the famous Jones polynomial. Other invariants obtained are more powerful than that of Jones. Powerful methods for completely analytical and non-perturbative computation of these knot and link invariants have been developed. In the process answers to some of the open problems in knot theory are obtained. From these invariants for unoriented and framed links in  $S^3$ , an invariant for any three-manifold can be easily constructed by exploiting the Lickorish-Wallace surgery presentation of three-manifolds. This invariant up to a normalization is the partition function of the Chern-Simons field theory. Even perturbative analysis of the Chern-Simons theories are rich in their mathematical structure; these provide a field theoretic interpretation of Vassiliev knot invariants. Not only in mathematics, Chern-Simons theories find important applications in three and four dimensional quantum gravity also.

---

\*Email: kaul@imsc.ernet.in

## 1 Introduction

Many a time advances in mathematics and physics have occurred hand in hand. Newton's theory of mechanics and developments of the techniques of calculus are a classical example of this phenomenon. Another example is the developments in differential geometry inspired by the Maxwell theory of electromagnetism and Einstein theory of general relativity. A recent glorious example is the developments of topological quantum field theories and their relevance to the study of geometry and topology of low dimensional manifolds.

The application of topological quantum field theories reflects the enormous interest generated both by mathematicians and field theoreticians in building a link between quantum physics through its path integral formulation on one hand and geometry and topology of low dimensional manifolds on the other. These are indeed deep links which are only now getting explored. It does appear that the properties of low dimensional manifolds can be very successfully unraveled by relating them to infinite dimensional manifolds of fields. This provides a powerful tool to study these manifolds notwithstanding the 'lack of mathematical rigour' in defining the functional integrals of quantum field theory. Indeed, an axiomatic formulation of topological quantum field theories has also been attempted.

Topological quantum field theories are independent of the metric of the curved manifold on which these are defined; the expectation value of the energy-momentum tensor is zero,  $\langle T_{\mu\nu} \rangle = 0$ . These possess no local propagating degrees of freedom; only degrees of freedom are topological. Operators of interest in such a theory are also metric independent.

To illustrate how ideas of quantum field theory can be used to study topology of low dimensional manifolds, we shall focus our attention here on recent important developments in Chern-Simons gauge field theory as a topological quantum field theory on a three-manifold. This theory provides a field theoretic framework for the study of knots and links in a given three manifold [1] - [5]. It was A.S. Schwarz who first conjectured [3] that the now famous Jones polynomial [6] may be related to Chern-Simons theory. E. Witten in his pioneering paper about ten years ago demonstrated this connection [2]. In addition, he set up a general field theoretic framework to study knots and links. Since then enormous effort has gone into developing an exact and explicit non-perturbative solution of this field theory. Many of the standard techniques of field theory find applications in these developments. The interplay between quantum field theory and knot theory has paid rich dividends in both directions. Many of the open problems in knot theory have found answers in the process.

Wilson loop operators are the topological operators of the Chern-Simons gauge field theory. Their vacuum expectation values are the topological invariants for knots and links which do not depend on the exact shape, location or form of the knots and links but reflect only their topological properties. The power of this framework is so deep that it allows us to study these invariants not only on simple manifold such as three-sphere  $S^3$  but also on any arbitrary three-manifold.

The knot and link invariants obtained from these field theories are also intimately related to the integrable vertex models in two dimensions [7, 5]. These invariants have also been approached in different mathematical frameworks. A quantum group approach to these polynomial invariants has been developed [8]. Last decade or so has seen enormous activity in these directions in algebraic topology.

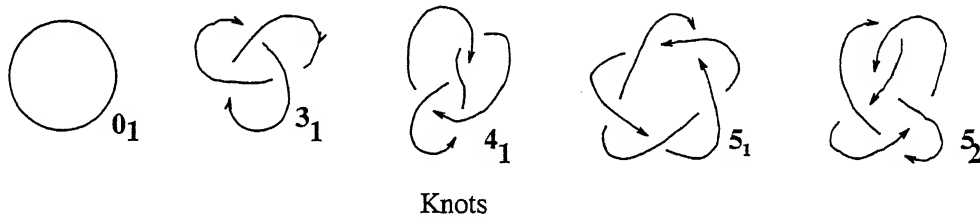
A mathematically important development is that these link invariants provide a method of obtaining a specific topological invariant for three-manifolds [2, 9] in terms of invariants for framed unoriented links in  $S^3$  [10, 5, 11]. In the following, we shall review these developments.

Not only in mathematics, Chern-Simons theory has also played a major role in quantum gravity. Three-dimensional gravity with a negative cosmological constant, itself a topological field theory, can be described by two copies of  $SU(2)$  Chern-Simons theory. Even in four dimensional gravity, Chern-Simons theories find application. For example, the boundary degrees of freedom of a black hole in four dimensions, are described by an  $SU(2)$  Chern-Simons field theory. This has allowed an exact calculation of quantum entropy of a non-rotating black hole. The formula so obtained for a Schwarzschild black hole, while agreeing with the Bekenstein-Hawking formula for large areas, goes beyond the semi-classical result.

Before explaining how a field theoretic framework for knots and links can be developed, let us start with a brief discussion of knots and links.

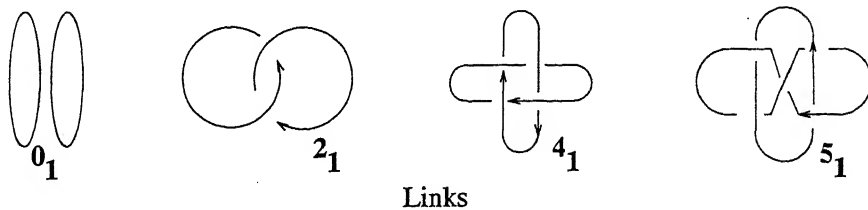
## 2 Knots and links: an elementary introduction

*What is knot?* A smooth non-intersecting closed curve in a three-manifold is a knot. Oriented closed curves are oriented knots. A string with its ends joined in the shape of a circle without any entanglements is a model for the simplest non-intersecting closed curve called *unknot*. With a given knot, we associate a *knot diagram* obtained by projecting the knot on to a plane with a minimum number of double points. In such a diagram over-crossings and under-crossings are to be clearly marked. The number of double points in a knot diagram is called its *crossing number*. A few simple knots with low crossing numbers are:



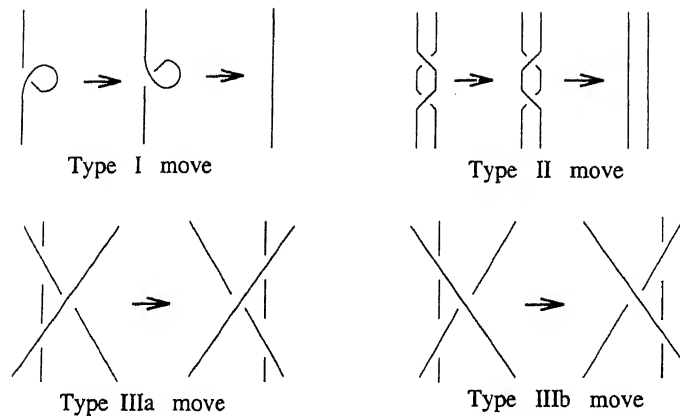
Clearly, for a given minimum number of crossings, there can be more than one type of topologically inequivalent knots. The number of knots increases rapidly with the crossing number. For crossing number 9, there are 49 knots (not distinguishing mirror reflections), for 10 there are 165 and for crossing number 11 we have 552 knots. For 13 crossings, there are more than 10,000 different knots.

*What is a link?* A collection of a number of oriented non-intersecting loops (knots) is an oriented link. A knot then is single component link. Links like knots can be represented by their two dimensional projection, *the link diagrams* with minimum number of double points, but with the over-crossings and under-crossings clearly marked. Examples of a few two-component links are:



To a topologist, length, thickness or precise shape of a knot are not of any interest. Two knots or links are to be identified if one can be made to go continuously into other by shrinking or stretching or wiggling without snapping the string. There is a minimal set of elementary rules which encode these qualitative notions more precisely. These are the three Reidemeister moves which do not change the topological type of a link:

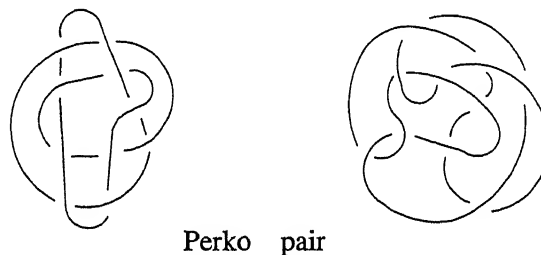




Invariance under all these three moves is called invariance under *ambient isotopy*. If a quantity is invariant under type II and III moves only, but not under type I moves, it is said to be a *regular isotopic invariant*.

The Reidemeister move III is of particular interest. It represents a defining relation for the generators of braids. In addition, it is a graphical representation of the Yang-Baxter relation of statistical mechanical models. These facts are not accidental but reflect a deep connection that knots and links have with braids and exactly solvable two-dimensional vertex models [7]. In fact this connection has been successfully exploited to obtain infinitely many new exactly solvable statistical mechanical models[5].

Though the Reidemeister rules are so simple, it is not an easy exercise in general to tell whether given two knots or links are topologically distinct or not. For example, it took nearly eighty years, since the time of knot tables of C.N. Little from the end of last century to the work of K.A. Perko in 1972, to recognize that the knots in the figure below are isotopically equivalent[12]:



Finding mathematical methods for distinguishing knots and links is indeed an important problem in knot theory. To this end, some definite invariants, called *link invariants*, are associated with the links. These are mathematical expressions which depend only on the isotopic type of the link and not on any of its particular representations. Some such invariants are in the form of polynomials. First polynomial invariant was discovered in late twenties by J.W. Alexander[13]. It took almost sixty more years before the next one was discovered by V.R.F. Jones[6]. The new invariant proved to be topologically more powerful than that of Alexander. For example, unlike Alexander polynomial, Jones polynomial does distinguish many mirror reflected knots. Soon after, a two variable generalization of Jones invariant was found [14]. Though two distinct Jones polynomials do represent two isotopically distinct knots, the converse is not always true. There are examples of distinct knots with same Jones polynomial. Still Jones work represents a leap forward in the developments of knot theory. What is impressive about the topological field theoretic description of knots is that it provides a whole variety of link invariants in a straight forward manner. Of these Jones one-variable polynomial and its two-variable generalization are the simplest examples.

Before starting a discussion of knots and links in terms of quantum field theory, let us make a few historical remarks about knots and links in physics.

**A few historical remarks:** Knots and links first captured the imagination of physicists when Lord Kelvin (William Thomas) introduced them as early as 1857 as fluid-mechanical models of atoms [15]. Reluctant to accept the prevailing notion of an infinitely rigid point-like atom, he thought of atoms as vortex-lines in a perfect homogeneous fluid, the *ether*. Different sorts of atoms were then to differ in accordance with the number of intersections of these vortex rings. “Stability” of the atoms in this theory thus is a reflection of the fact that knots do preserve their essential knottedness during their movement. Indeed Lord Kelvin would have wanted to develop a new theory of gasses, theory of elastic solids and liquids based on the dynamics of these vortex atoms – a programme he did not complete nor was considered by later day physicists worth while in this context. However, a new area, *knot theory*, of mathematics was born.

Two contemporary Scottish physicists, J.C. Maxwell and P.G. Tait did find Lord Kelvin’s hypothesis attractive enough. Tait had hoped to explain the position of lines in the spectrum of a chemical element from the knot type representing it. Thus, it was natural for him to attempt the formidable task of classifying knots in 3-space. For this he needed some measure of complexity of a knot. Thus the concept of the *degree of knottedness* was introduced. This is what we nowadays call crossing number of a knot, a notion already defined above. Tait with this notion of crossing number, produced the first knot tables, listing knots in order of their increasing knottedness. If atoms had been really knots, we would have been studying these tables instead of the period table of chemical elements in our schools.

Since the pioneering work of these physicists, knot theory was solely investigated by mathematicians till about ten years ago when physicists came back to it through quantum field theories. This brings us to modern field theoretic interpretation of knots in three dimensions.

### 3 Abelian Chern-Simons field theory and knots and links

In a field theory, the properties of a system of infinitely many oscillators are represented collectively by a field,  $\phi(x)$  defined over all the space though the space label  $x$ . An action functional is prescribed for these fields. For example, for a one-component scalar field  $\phi(x)$ , say in three dimensional flat Euclidean space  $R^3$ , the action functional may be taken to be:

$$S[\phi] = \frac{1}{2} \int d^3x \delta^{\mu\nu} \partial_\mu \phi(x) \partial_\nu \phi(x) ,$$

where  $\mu, \nu = 1, 2, 3$  are space indices and for  $R^3$ , the metric is flat  $\delta^{\mu\nu} = \text{dia}(1, 1, 1)$ . For a theory defined over a general curved three-manifold endowed with a metric  $g_{\mu\nu}$  (and its inverse  $g^{\mu\nu}$ ), this action generalizes to:

$$S[\phi] = \frac{1}{2} \int d^3x \sqrt{g(x)} g^{\mu\nu}(x) \partial_\mu \phi(x) \partial_\nu \phi(x) ,$$

where  $g(x) = \det g_{\mu\nu}$ .

Similarly for a vector field  $A_\mu(x)$ , the gauge field of the Maxwell theory in three dimensions, we write the action functional as:

$$S[A_\mu] = \frac{1}{4} \int d^3x \sqrt{g(x)} g^{\mu\alpha}(x) g^{\nu\beta}(x) \left[ \partial_\mu A_\nu(x) - \partial_\nu A_\mu(x) \right] \left[ \partial_\alpha A_\beta(x) - \partial_\beta A_\alpha(x) \right]$$

Both these actions above are invariant under general coordinate transformations.

Quantum field theories normally studied, like the examples above, depend on the metric  $g_{\mu\nu}$  of the three-manifold in which the theory is defined. The metric describes the geometric properties, such as distances, curvature etc. But, here we are interested in attempting a field theoretic description of knots and links in such a way that only their topological properties are represented. Their size, exact shape, location etc are not of our concern. The topological properties, unlike these, do not depend on the metric. Thus we are seeking a field theory which is independent of the metric. Such theories are called *topological field theories*. A simple example of metric independent

field theory is the Chern-Simons gauge theory. Its action in the Abelian version is given (with convenient normalization) by:

$$kS[A_\mu] = -\frac{k}{8\pi} \int_{S^3} d^3x \epsilon^{\mu\nu\alpha} A_\mu(x) \partial_\nu A_\alpha(x) \quad (1)$$

where  $\epsilon^{\mu\nu\alpha}$  is a completely anti-symmetric contravariant three-tensor density whose only nonzero component is  $\epsilon^{123} = 1$ . For definiteness, we shall discuss this theory in a three-manifold  $S^3$ . Clearly this action is independent of the metric. Also it is invariant under general coordinate transformations. Like the Maxwell theory, this theory exhibits a gauge invariance.

The quantum version of this theory is described by the functional integral representing the partition function:

$$Z = \int [dA] e^{ikS} \quad (2)$$

and for metric independent gauge invariant functionals  $W[A_\mu]$  of the gauge field  $A_\mu(x)$ , we have the functional averages (vacuum expectation values of the associated operators):

$$\langle W \rangle = Z^{-1} \int [dA] W e^{ikS} \quad (3)$$

Though the action and gauge invariant functionals  $W$  do not depend on the metric, there are potential sources which can introduce metric dependence in these functional averages. The functional integration may be thought of to be done by discretizing the space into a mesh. Infinitely many ordinary integrals over  $A_\mu(x)$  at each point  $x$  of the mesh are to be done and finally the limit of mesh size going to zero is taken in some well defined manner. This is the usual way we understand these infinite dimensional integrals. Further, there is a gauge invariance in the theory, which like other gauge theories needs to be fixed by a choice of gauge. Both the choice of mesh as well as gauge fixing condition are generically metric dependent. Thus the gauge fixed measure of integration  $[dA_\mu(x)]$  in a field theory defined on a curved space, in general depends on the metric. However, despite these, it can be shown that various metric dependence so conspire in this topological theory that they cancel out without spoiling the metric independence of the functional averages [16].

Now let us give an explicit form of a topological operator  $W$  in this Abelian Chern-Simons theory. Consider a link  $L$  made up of knots  $K_1, K_2, \dots, K_s$ . Wilson knot operator for each these knots  $K_\ell$  is given by  $\exp[i n_\ell \oint_{K_\ell} dx^\mu A_\mu(x)]$  where  $n_\ell$  is an integer measuring the charge on the loop. Clearly these are independent of the metric. Then the Wilson link the operator is product of all such knot operators:

$$W[L] = \prod_{\ell=1}^s \exp\left[i n_\ell \oint_{K_\ell} dx^\mu A_\mu(x)\right] \quad (4)$$

If we expand the exponential here, the expectation value  $\langle W[L] \rangle$  is given by the expectation values of the various terms in this expansion. This is a non-interacting theory, all these expectation values are given in terms of the "two-loop" expectation values only:

$$\left\langle \oint_{K_\ell} dx^\mu A_\mu(x) \oint_{K_m} dy^\nu A_\nu(y) \right\rangle, \quad K_\ell \neq K_m; \quad \text{and} \quad \left\langle \oint_K dx^\mu A_\mu(x) \oint_K dy^\nu A_\nu(y) \right\rangle \quad (5)$$

Here in the first expression the two loops are distinct in contrast to the second expression where both the loop integrals are along the same knot. Clearly, these expressions can be easily evaluated in terms of the two-point correlator  $\langle A_\mu(x) A_\nu(y) \rangle$ . To do this, we can locally identify the region containing our link with  $R^3$  so that we can use the flat metric  $g_{\mu\nu} = \delta_{\mu\nu}$  in this region. Then  $x^\mu$  and  $y^\nu$  are the Euclidean flat coordinates along the two knots  $K_\ell$  and  $K_m$  respectively. This allows us to do away with the complications connected with the curved nature of the three-manifold  $S^3$ ; we can do all our calculations in flat Euclidean space without loss of generality. Elementary field

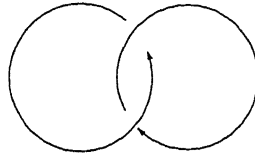
theory allows us to read off the flat space two-point correlator from the action (subject to a gauge condition, which we choose to be the covariant Lorentz gauge  $\delta^{\mu\nu} \partial_\mu A_\nu = 0$ ):

$$\langle A_\mu(x) A_\nu(y) \rangle = \frac{i}{k} \epsilon_{\mu\nu\alpha} \frac{(x-y)^\alpha}{|x-y|^3}$$

so that

$$\begin{aligned} \langle \oint_{K_\ell} dx^\mu A_\mu(x) \oint_{K_m} dy^\nu A_\nu(y) \rangle &= \frac{4\pi i}{k} \mathcal{L}(K_\ell, K_m) \\ \text{where} \quad \mathcal{L}(K_\ell, K_m) &= \frac{1}{4\pi} \oint_{K_\ell} dx^\mu \oint_{K_m} dy^\nu \epsilon_{\mu\nu\alpha} \frac{(x-y)^\alpha}{|x-y|^3}. \end{aligned} \quad (6)$$

This double loop integral over two distinct knots ( $K_\ell \neq K_m$ ) is a well known topological invariant, called *Gauss linking number* of the two closed curves. It measures the number of times one knot  $K_\ell$  goes through the other knot  $K_m$ . Clearly, linking number of two knots is an integer. For example, for the right-handed Hopf link  $H_+$ ,



Right-handed Hopf link  $H_+$

its value is +1. Its value for the mirror reflection of this link (left-handed Hopf) is -1. Linking number does not depend on the exact location of the two knots, nor on their size or shape. It depends only on their topological relationship with each other. This invariant has a physical interpretation due to Maxwell – in electrodynamics, it represents the work done to move a magnetic monopole around one knot in three-space while an electric current runs through the other knot.

The Abelian Chern-Simons theory also provides a representation for yet another simple topological quantity associated with an individual knot called its *self-linking number* and also some times *framing number* or simply *framing*. This is related to the second expectation value given in (5) where the two loop integrals are over the same knot. This expectation value is to be evaluated through a limiting procedure: To a knot  $K$  parametrized by  $x^\mu(s)$  ( $0 \leq s \leq L$ ) along the length of the knot by the parameter  $s$ , associate another closed curve  $K_f$ , called its *frame*, given by coordinates  $y^\mu = x^\mu(s) + \epsilon n^\mu(s)$  where  $\epsilon$  is a small parameter and  $n^\mu(s)$  is a unit vector field normal (principal normal) to the curve at  $s$ . That is,  $K_f$  is the curve  $K$  displaced along the normal by a small amount. Then the linking number of the curve  $K$  and its frame  $K_f$  is called the self-linking number  $\mathcal{SL}(K)$  of the knot:

$$\begin{aligned} \langle \oint_K dx^\mu A_\mu(x) \oint_{K_f} dy^\nu A_\nu(y) \rangle &= \lim_{\epsilon \rightarrow 0} \langle \oint_K dx^\mu A_\mu(x) \oint_{K_f} dy^\nu A_\nu(y) \rangle \\ &= \frac{4\pi i}{k} \mathcal{L}(K, K_f) = \frac{4\pi i}{k} \mathcal{SL}(K) \end{aligned} \quad (7)$$

This self-linking number is independent of the parameter  $\epsilon$  and can easily be shown to obey the following important theorem, first proven by G. Calugareanu almost forty years ago [17]:

**Calugareanu theorem:** *The self-linking number of a knot is the sum of its twist and writhe numbers:*

$$\mathcal{SL}(K) = T(K) + w(K) \quad (8)$$

$$T(K) = \frac{1}{2\pi} \int_K ds \epsilon_{\mu\nu\alpha} \frac{dx^\mu}{ds} n^\nu \frac{dn^\alpha}{ds}, \quad w(K) = \frac{1}{4\pi} \int_K ds \int_K dt \epsilon_{\mu\nu\alpha} e^\mu \frac{de^\nu}{ds} \frac{de^\alpha}{dt}$$

where the vector field  $e^\mu$  is given by

$$e^\mu(s, t) = \frac{y^\mu(t) - x^\mu(s)}{|y(t) - x(s)|}$$

is a map  $K \otimes K \mapsto S^2$  and  $n^\mu(s)$  is the normal vector field along the length of the curve  $K$  ( $x^\mu(s), 0 \leq s \leq L$ ). The quantities  $T(K)$  and  $w(K)$  represent well defined *geometric* properties of the knot.  $T(K)$  represents the *twist* in the knot  $K$  with reference to its frame  $K_f$  and  $w(K)$  is the amount of *writhe* or *coiling* of the knot. Clearly, the twist number and writhe number are not necessarily integers nor are they ambient isotopic invariants. But their sum, the self-linking number, is indeed an integer and also an ambient isotopic invariant. This theorem can be easily appreciated if we recall that stretching a coiled up telephone cord reduces its coils but increases its twist and loosening of a twisted cord coils it up. The amount of coils lost (or gained) is exactly the same as the amount by which the twisting is gained (or lost) so that their sum is always unchanged. This theorem of Calugareanu when applied to circular ribbon (which can be thought of as a framed closed curve) has been put to good use in the study of the properties of circular polymers and circular DNA [18].

Notice that the self-linking number does carry dependence on the frame. The mathematical concept of framing of a knot is intimately connected to the concept of regularization in field theory. In order to avoid the coincidence singularity in the two-point correlator  $\lim_{x \rightarrow y} \langle A_\mu(x) A_\nu(y) \rangle$ , we need to regularize it, say by point-splitting. Evaluating, 'two-loop' correlator of Eqn.(5), where the two loops are same, we face this same divergence, which, through framing, has been resolved by 'loop-splitting'. Ordinarily, quantities in field theory do depend on the regularization. Like-wise the self-linking number here depends on the framing. But all those framing curves enveloping around the knot, which can be continuously deformed into each other without snapping the knot, form a topological class for which the self-linking number does not change. In field theory language, framing provides a *topological regularization*.

Now collecting all these pieces of information, the expectation value of the Wilson link operator for a link  $L = (K_1, K_2, \dots, K_s)$  in the Abelian Chern-Simons theory on  $S^3$  can be written down in terms of the linking and self-linking (framing) numbers as:

$$\langle W[L] \rangle = \exp \left\{ -\frac{2\pi i}{k} \left[ \sum_{\ell} n_{\ell}^2 \mathcal{SL}(K_{\ell}) + \sum_{\ell \neq m} n_{\ell} n_m \mathcal{L}(K_{\ell}, K_m) \right] \right\} \quad (9)$$

Thus, we have indicated here how this simple field theory does indeed, through expectation values of Wilson link operators, provide a field theoretic interpretation of some of the topological invariants, linking number and self-linking number of knots and links. Non-Abelian Chern-Simons theories are much richer in their structure; these capture even more complex topological properties of knots and links.

## 4 Non-Abelian Chern-Simons field theory as a description of knots and links

A non-Abelian Chern-Simons theory, instead of being a gauge theory of one vector field, involves, say for gauge group  $SU(2)$ , three such fields,  $A_\mu^a$  ( $a = 1, 2, 3$ ). These three are collectively written as a matrix valued vector field  $A_\mu = A_\mu^a \frac{\sigma^a}{2i}$ , where anti-hermetian matrices  $\frac{\sigma^a}{2i}$  are the generators of the group  $SU(2)$ . Action functional defined in a three-manifold, say  $S^3$ , is given by:

$$kS = \frac{k}{4\pi} \int_{S^3} d^3x \epsilon^{\mu\nu\alpha} \text{tr} \left[ A_\mu(x) \partial_\nu A_\alpha(x) + \frac{2}{3} A_\mu(x) A_\nu(x) A_\alpha(x) \right] \quad (10)$$

Like Abelian Chern-Simons theory, this action has no metric dependence. Besides a gauge invariance, it is also invariant under general coordinate transformations.

The topological operators are the Wilson loop (knot) operators defined as

$$W_j[K] = \text{tr}_j \text{Pexp} \oint_K dx^\mu A_\mu^a T_j^a \quad (11)$$

for an oriented knot  $K$  carrying spin  $j$  representation reflected by the associated representation matrices  $T_j^a$  ( $a = 1, 2, 3$ ). The symbol  $P$  stands for path ordering of the exponential. This is done by breaking the length of the knot  $K$  into infinitesimal intervals of size  $dx_m^\mu$  around the points labeled by the coordinates  $x_m^\mu$  along the knot. Then path ordered exponential is:

$$P \exp \oint_K dx^\mu A_\mu^a T_j^a = \prod_m [1 + dx_m^\mu A_\mu^a(x_m) T_j^a]$$

For a link  $L$  made up of oriented component knots  $K_1, K_2, \dots, K_s$  carrying spin  $j_1, j_2, \dots, j_s$  representations respectively, we have the Wilson link operator defined as

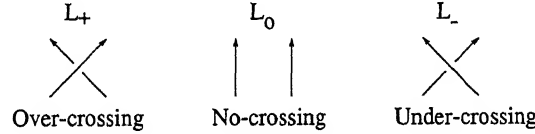
$$W_{j_1 j_2 \dots j_s}[L] = \prod_{\ell=1}^s W_{j_\ell}[K_\ell] \quad (12)$$

We are interested in the functional averages of these operators:

$$V_{j_1 j_2 \dots j_s}[L] = Z^{-1} \int [dA] W_{j_1 j_2 \dots j_s}[L] e^{ikS}, \quad \text{where } Z = \int [dA] e^{ikS} \quad (13)$$

Here the integrands in the functional integrals are metric independent. So is the measure [16]. Therefore, these expectation values depend only on the isotopy type of the oriented link  $L$  and the set of representations  $j_1, j_2, \dots, j_s$  associated with component knots.

These expectation values can be obtained non-perturbatively. For example, for knots and links carrying only the spin 1/2 representations, Witten has shown that the link invariants (expectation values of the associated Wilson link operators) satisfy a simple relation. This relation is given for three link diagrams which are identical every where except for one crossing where they differ in that it is an over-crossing ( $L_+$ ), or no-crossing ( $L_0$ ) or an under-crossing ( $L_-$ ) as shown in the figure below:

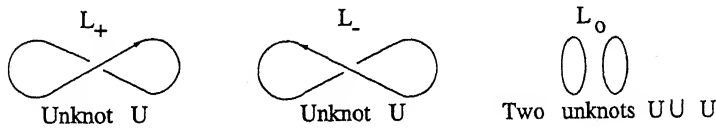


Then the invariant for such links are related as:

$$q V_{1/2}[L_+] - q^{-1} V_{1/2}[L_-] = (q^{1/2} - q^{-1/2}) V_{1/2}[L_0] \quad (14)$$

where  $q$  is a root of unity related to the Chern-Simons coupling  $k$  through the relation  $q = \exp[2\pi i/(k+2)]$ . This is precisely the well known generating skein relation for the Jones polynomials. Indeed  $V_{1/2}[L]$ , which is the expectation value of the Wilson link operator where every component knot carries the doublet spin 1/2 representation of  $SU(2)$ , is the one-variable Jones polynomial.

The above skein relation is powerful enough that it recursively yields Jones polynomial for any arbitrary link. For example consider following three link diagrams:

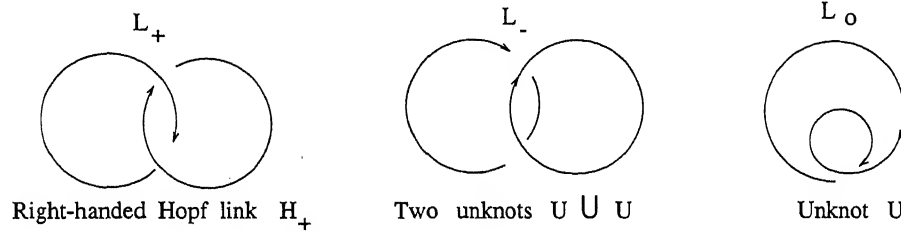


We use an important factorization property of these invariant: *the link invariant of two distant (disjoint) links (that is, with no mutual entanglement) is simply the product of invariants for the individual links*. That is, for the link  $L_0$  above,  $V_{1/2}[U \cup U] = (V_{1/2}[U])^2$ , where symbol  $U$  represents the unknot. Then use of the skein relation yields:

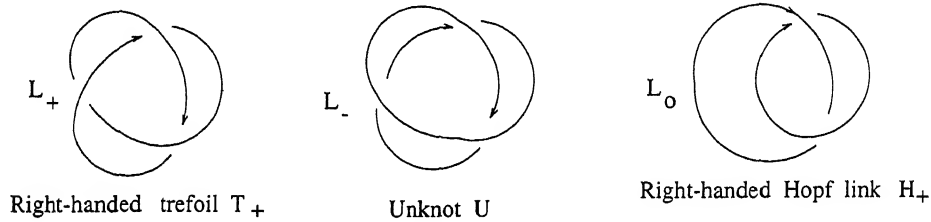
$$q V_{1/2}[U] - q^{-1} V_{1/2}[U] = (q^{1/2} - q^{-1/2}) (V_{1/2}[U])^2$$

so that spin 1/2 invariant for an unknot is given by:  $V_{1/2}[U] = q^{1/2} + q^{-1/2}$ .

Next apply the skein relation to three links, where the  $L_+$  is the right-handed Hopf link,  $L_-$  is simply the union of two (unlinked) unknots and  $L_0$  is an unknot:



This yields, the invariant for the right-handed Hopf link  $H_+$  as:  $V_{1/2}[H_+] = 1 + q^{-1} + q^{-2} + q^{-3}$ . Now use recursion relation for the three links:

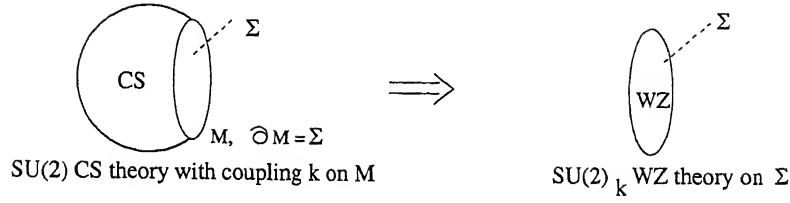


where  $L_+$  is a right-handed trefoil ( $T_+$ ),  $L_-$  is an unknot and  $L_0$  is a right-handed Hopf  $H_+$ . This gives us the invariant for the trefoil knot as  $V_{1/2}[T_+] = q^{-1/2} + q^{-3/2} + q^{-5/2} - q^{-9/2}$ . This way invariant for any arbitrary link can be recursively obtained.

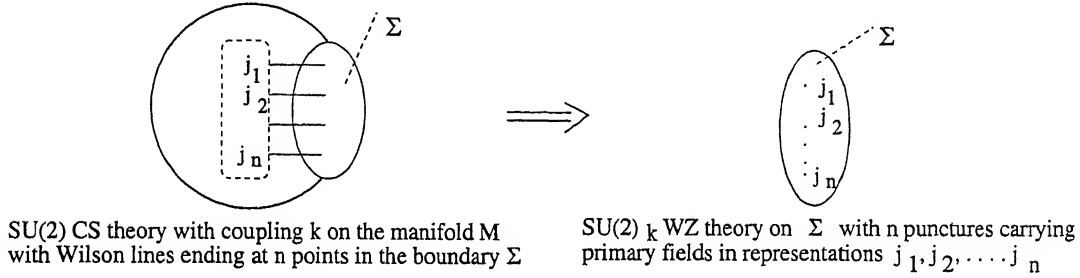
Jones polynomial is in fact the simplest of the examples of a whole host new link invariants that emerge naturally from this field theory. More general invariants are the expectation values of Wilson link operators with arbitrary spin representations placed on the knots. The formalism does also allow for placing different representations on each of the component knots. This leads to so-called *coloured polynomial invariants*. Besides, instead of the gauge group  $SU(2)$ , Chern-Simons theory based on any other semi-simple group can be used. These then yield even richer spectrum of the new invariants.

While Jones polynomial can be obtained by recursive use of the skein relation, other more general invariants (for spin representations  $j = 1, 3/2, \dots$ ) can *not* be obtained in this manner. Of course there are generalizations of the skein relations for an arbitrary spin invariants. But these do not possess recursively complete solutions (except for spin 1/2 case above). Therefore methods had to be developed to obtain expectation values of Wilson operators with arbitrary representations living on the component knots of a link. One such method in its complete manifestations has been presented in ref [4]. This allows us to present a complete and explicit solution of the Chern-Simons theory. This is a non-perturbative method which, generalizing the formalism set up by Witten, makes use of two ingredients, one from quantum field theory and other from mathematics of braids:

(i) *Field theoretic input:* Chern-Simons theory on a three-manifold with boundary is essentially characterized by a corresponding two dimensional Wess-Zumino conformal field theory on that boundary[2]:



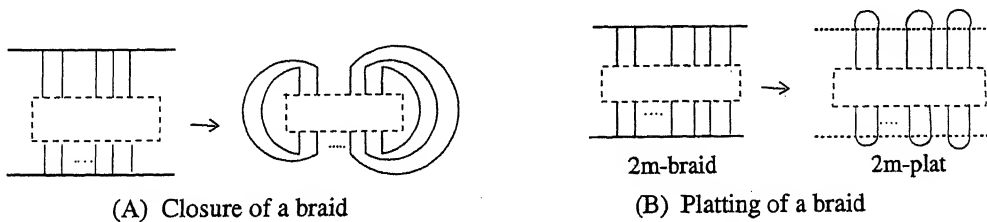
And Chern-Simons functional average for Wilson lines ending at  $n$  points in the boundary is described by the associated Wess-Zumino theory on the boundary with  $n$  punctures carrying the representations of the free Wilson lines:



The Chern-Simons functional integral can be represented [2] by a vector in the Hilbert space  $\mathcal{H}$  associated with the space of  $n$ -point correlator of the Wess-Zumino conformal field theory on the boundary  $\Sigma$ . In fact, these correlators provide a basis for this boundary Hilbert space. There are more than one possible basis. These different bases are related by duality of the correlators of the conformal field theory[4].

(ii) *Mathematical input:* The second ingredient used is the close connection knots and links have with braids. An  $n$ -braid is a collection of non-intersecting strands connecting  $n$  points on a horizontal plane to  $n$  points on another horizontal plane directly below the first set of  $n$  points. The strands are not allowed to go back upwards at any point in their travel. The braid may be projected onto a plane with the two horizontal planes collapsing to two parallel rigid rods. The over-crossings and under-crossings of the strands are to be clearly marked. When all the strands are identical, we have ordinary braids. The theory of such braids, first developed by Artin, is well studied. These braids form a group. However, we may wish to orient the individual strands and further distinguish them by putting different colours on them. These different colours are represented by different  $SU(2)$  spins. These braids, unlike braids made from unoriented identical strands, have a more general structure than a group. These instead form a groupoid. The necessary aspects of the theory of such braids have been presented in ref.[4]

One way of relating the braids to knots and links is through closure of braids. We obtain the closure of a braid by connecting the ends of the first, second, third, ..... strands from above to the ends of the respective first, second, third, ..... strands from below as shown in (A):



There is a theorem by Alexander[19] which states that *any knot or link can be obtained as closure of a braid*. This construction of a knot or link is not unique.

There is another construction associated with braids which relates them to knots and links.



This is called platting. Consider a  $2m$ -braid, with pairwise adjacent strands carrying the same colour and opposite orientations. Then connect the  $(2i - 1)$ th strand with  $(2i)$ th from above as well as from below. This yields the plat of the given braid as shown in (B) above. There is a theorem due to Birman[20] which relates plats to links. This states that *a coloured-oriented link can be represented (though not uniquely) by the plat of an oriented-coloured  $2m$ -braid.*

Use of these two inputs, namely relation of Chern-Simons theory to the boundary Wess-Zumino conformal field theory and presentation of knots and links as closures or plats of braids leads to an explicit, complete and non-perturbative solution of the Chern-Simons theory. Conformal field theory on associated boundary gives matrix representations for braids and platting or closing of a braid corresponds to taking a specific matrix element of these braid representations. This then yields the expectation value of the Wilson link operator associated with that link. For example this invariant for an unknot  $U$  carrying spin  $j$  representation turns out to be:

$$V_j[U] = [2j + 1] \quad \text{where} \quad [x] = \frac{q^{x/2} - q^{-x/2}}{q^{1/2} - q^{-1/2}}$$

The square bracket indicates a  $q$ -number. Jones polynomial above corresponds to spin  $j = 1/2$ . And for a right-handed trefoil  $T_+$ , the invariant turns out to be:

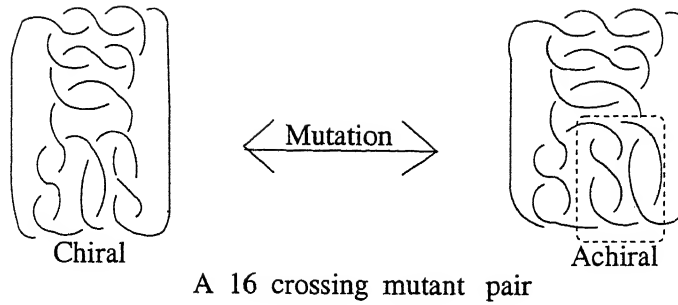
$$V_j[T_+] = \sum_{m=0,1,2,\dots,\min(2j,k-2j)} [2m + 1] (-)^{2j+m} q^{-6C_j + \frac{3}{2}C_m}$$

For  $j = 1/2$ , this agrees with the polynomial obtained above by using the skein relation.

The link invariants calculated from the field theory depend on the regularization used to define the coincident loop correlators, that is, the framing of the knots. The invariants above have been obtained in a specific framing called *standard framing*. In particular, the skein relation for spin  $1/2$  invariants given above is in this framing. In this framing, the self-linking (framing) number of every knot is zero. The invariants so obtained are unchanged under all the three Reidemeister moves. That is, this yields ambient isotopic invariants. There is another framing choice which has been of special interest. In this case, the frame is thought to be just vertically above the two dimensional projection of the knot. In this framing, known as *vertical framing*, Reidemeister moves II and III do leave the link invariants unchanged, but Reidemeister move I changes them.

The general framework developed provides a powerful method of calculating knot and link invariants. This has in the process also provided answers to some of the open problems of knot theory. For example, one such problem is to find polynomial invariants which would discriminate between two chiralities of a given knot. The invariants for the mirror reflected knots are given by simple complex conjugation. Up to ten crossing number, there are six chiral knots,  $9_{42}$ ,  $10_{48}$ ,  $10_{71}$ ,  $10_{91}$ ,  $10_{104}$  and  $10_{125}$  (as listed in the knot tables in Rolfsen's book [21]) which are not distinguished from their mirror images by spin  $1/2$  (Jones) polynomials. Spin one (Kauffman/Akutsu-Wadati) polynomials do detect the chirality of four of them, namely  $10_{48}$ ,  $10_{91}$ ,  $10_{104}$  and  $10_{125}$ . But for  $9_{42}$  and  $10_{71}$  both Jones and Kauffman polynomials are not changed under chirality transformation ( $q \rightarrow q^{-1}$ ). However, the new spin  $3/2$  invariants are powerful enough to distinguish these knots from their mirror images[22].

Another problem of knot theory that has been provided with an answer is to do with so called *mutant* knots. A mutant of a knot or link is obtained in the following way: isolate a portion of the knot in such a way that it has two strands going into and two strands leaving from it. Scoop it out and rotate it through  $\pi$  about any of three orthogonal axes (rotations about only two of these are really independent). Glue it back after, if necessary, changing the orientations on the strands to match the free ends of strands of the rest of the knot to which the free ends of the rotated portion are glued. This yields a mutant of the original knot. It has been possible to prove that polynomial invariants obtained from a Chern-Simons theory based on *any arbitrary non-Abelian gauge group* do not distinguish isotopically inequivalent mutant knots[23]. As an example consider the following sixteen crossing mutant knots:



The two knots are related by a mutation of the portion indicated by dashed enclosure. Like all other mutants, the invariants obtained from any non-Abelian Chern-Simons theory for them are identical. What is of particular interest about this pair is that one of them is chiral, other is not. This then yields an example of a chiral knot whose chirality can not be detected by any of these invariants.

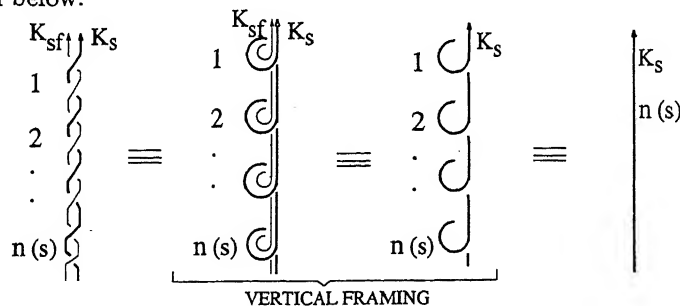
The general framework developed to study knots and links is also applicable to another set of gauge invariant operators called graphs. For  $SU(2)$  Chern-Simons theory, these are the graphs containing vertices with three legs. The edges of the graph between vertices carry Wilson line operators. More general gauge invariant operators which include links attached to the edges of graphs can also be evaluated in this framework.

## 5 Three-manifold invariants

The invariants of knots and links in  $S^3$  obtained from the Chern-Simons theory can be used to construct a special three-manifold invariant [2, 9, 10, 5]. This provides an important tool to study topological properties of three-manifolds. Starting step in this construction is a theorem due to Lickorish and Wallace [24, 21]:

**Fundamental theorem of Lickorish and Wallace:** *Every closed, orientable, connected three-manifold,  $M^3$  can be obtained by surgery on an unoriented framed knot or link  $[L, f]$  in  $S^3$ .*

As described earlier, the framing  $f$  of a link  $L$  is defined by associating with every component knot  $K_s$  of the link an accompanying closed curve  $K_{sf}$  parallel to the knot and winding  $n(s)$  times in the right-handed direction. That is, the linking number  $lk(K_s, K_{sf})$  of the component knot and its frame (self-linking number of the knot  $K_s$ ) is  $n(s)$ . For the construction of three-manifold invariants, we use vertical framing where where the frame is thought to be just vertically above the two dimensional projection of the knot as shown below. This is sometimes indicated by putting  $n(s)$  writhes in the strand making the knot or even by just simply writing the integer  $n(s)$  next to the knot as shown below:



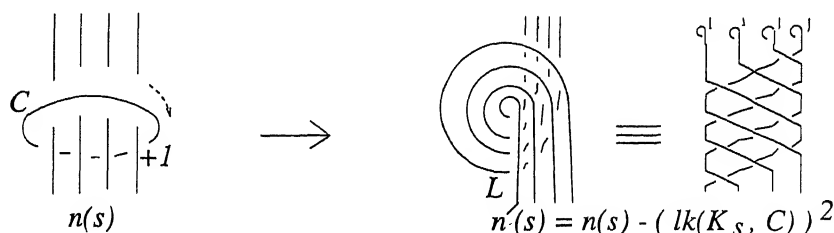
Next the surgery on a framed link  $[L, f]$  made of component knots  $K_1, K_2, \dots, K_r$  with framing  $f = (n(1), n(2), \dots, n(r))$  in  $S^3$  is performed in the following manner. Remove a small open solid torus neighbourhood  $N_s$  of each component knot  $K_s$ , disjoint from all other such open tubular neighbourhoods associated with other component knots. In the manifold left behind  $S^3 - (N_1 \cup N_2 \cup \dots \cup N_r)$ , there are  $r$  toral boundaries. On each such boundary, consider a simple closed curve (the frame) going  $n(s)$  times along the meridian and once along the longitude of the associated

knot  $K_s$ . Now do a modular transformation on such a toral boundary such that the framing curve bounds a disc. Glue back the solid tori into the gaps. This yields a new manifold  $M^3$ . The theorem of Lickorish and Wallace assures us that every closed, orientable, connected three-manifold can be constructed in this way.

This construction of three-manifolds by surgery is not unique: surgery on more than one framed link can yield homeomorphic manifolds. But the rules of equivalence of framed links in  $S^3$  which yield the same three-manifold on surgery are known. These rules are known as Kirby moves[25].

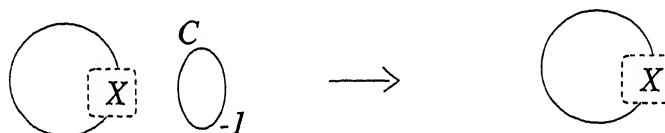
**Kirby calculus on framed links in  $S^3$ :** Following two elementary moves (and their inverses) generate Kirby calculus:

*Move I.* For a number of unlinked strands belonging to the component knots  $K_s$  with framing  $n(s)$  going through an unknotted circle  $C$  with framing  $+1$ , the unknotted circle can be removed after making a complete clockwise twist from below in the disc enclosed by the circle  $C$ :



In the process, in addition to introducing new crossings, the framing of the various resultant component knots,  $K'_s$  to which the affected strands belong, change from  $n(s)$  to  $n'(s) = n(s) - (lk(K_s, C))^2$ .

*Move II.* Drop a disjoint unknotted circle with framing  $-1$  without any change in the rest of the link:



Thus Lickorish-Wallace theorem and equivalence of surgery under Kirby moves reduces the theory of closed, orientable, connected three-manifolds to the theory of framed unoriented links via a one-to-one correspondence:

$$\left( \begin{array}{l} \text{Framed links in } S^3 \text{ modulo} \\ \text{equivalence under Kirby moves} \end{array} \right) \leftrightarrow \left( \begin{array}{l} \text{Closed, orientable, connected three-} \\ \text{manifolds modulo homeomorphisms} \end{array} \right)$$

This consequently allows us to characterize three-manifolds by the invariants of the associated unoriented framed knots and links obtained from the Chern-Simons theory in  $S^3$ . This can be done by constructing an appropriate combination of the invariants of the framed links which is unchanged under Kirby moves:

$$\left( \begin{array}{l} \text{Invariants of a framed unoriented link} \\ \text{which do not change under Kirby moves} \end{array} \right) = \left( \begin{array}{l} \text{Invariants of associated} \\ \text{three-manifold} \end{array} \right)$$

One such invariant has been constructed in ref [5]. It is given in terms of invariants for *un-oriented* links obtained from  $SU(2)$  Chern-Simons theory. The link invariants discussed in Sec.4 above are obtained in standard framing. These are sensitive to the relative orientations of the component knots. Here we shall use invariants for unoriented links in vertical framing. But, unlike the invariants in standard framing which exhibit ambient isotopic invariance, those obtained in vertical framing have only regular isotopic invariance. That is, in standard framing, a writhe can be stretched (a Reidemeister move I) without affecting the link invariant, in vertical framing this

is not so. The link invariant gets changed by a phase when a writhe is smoothed out as:

$$\text{R} \bigcirc \text{j} = q^{C_j} \bigcap^j, \quad \text{and} \quad \text{j} \bigcirc \text{L} = q^{-C_j} \bigcap^j$$

where we have represented the link invariant by the affected portion of the link. Thus, in vertical framing, invariant for an unknot with self-linking (framing) number +1 or -1 is related to the invariant for an unknot with zero self-linking number as:

$$V_j \left[ \text{R} \bigcirc^{+1} \right] = q^{C_j} V_j \left[ \bigcirc^0 \right] = q^{C_j} [2j+1],$$

$$\text{and } V_j \left[ \text{L} \bigcirc^{-1} \right] = q^{-C_j} V_j \left[ \bigcirc^0 \right] = q^{-C_j} [2j+1].$$

In this framing, each right-(left-) handed crossing in a knot introduces a self-linking number +1 (-1). For a right-handed trefoil (self-linking number = 3), the invariant in this framing turns out to be:

$$V_j[T_+] = \sum_{m=0,1,\dots,\min(2j,k-2j)} [2m+1] (-)^m q^{-3C_j+\frac{3}{2}C_m}$$

Three-manifold invariant is constructed from these link invariants in vertical framing. It has been shown that[5]: For a framed link  $[L, f]$  with component knots,  $K_1, K_2, \dots, K_r$  and their framings respectively as  $n(1), n(2), \dots, n(r)$ , the quantity

$$\hat{F}[L, f] = \alpha^{-\sigma[L, f]} \sum_{\{j_i\}} \mu_{j_1} \mu_{j_2} \dots \mu_{j_r} V[L; n(1), n(2), \dots, n(r); j_1, j_2, \dots, j_r] \quad (15)$$

constructed from invariants  $V$  of the unoriented framed link in vertical framing, is an invariant of the associated three-manifold obtained by surgery on that link. Here the coefficients  $\mu_\ell$  are given by

$$\mu_\ell = S_{0\ell}, \quad \text{where} \quad S_{j\ell} = \sqrt{\frac{2}{k+2}} \sin \frac{\pi(2j+1)(2\ell+1)}{k+2}.$$

and  $\alpha = \exp 3\pi i k / [4(k+2)]$ , and  $\sigma[L, f]$  is the signature of the linking matrix  $W[L, f]$ :  $\sigma[L, f] = (\text{no. of +ve eigenvalues of } W) - (\text{no. of -ve eigenvalues of } W)$ . The off diagonal elements of the linking matrix  $(W[L, f])_{ij}$  are given by linking number  $lk(K_i, K_j)$  for the distinct knots ( $i \neq j$ ) and diagonal elements ( $i = j$ ) are the self-linking number (frame number) of the knot  $K_i$ :  $(W[L, f])_{ii} = \mathcal{SL}(K_i) = n_i$ .

It can be directly verified that this three-manifold invariant (15) is unchanged under Kirby moves I and II.

*Explicit examples:* Now computation of this invariant for various three-manifolds is rather straight forward. We present its value for a few three-manifold. The surgery descriptions of manifolds  $S^3$ ,  $S^2 \times S^1$  and  $RP^3$  are given by an unknot with framing +1, 0 and +2 respectively. As indicated above the invariant for an unknot with zero framing carrying spin  $j$  representation is  $[2j+1] = S_{0j}/S_{00}$ , where the square bracket represents the  $q$ -number. Thus the invariant for  $S^3$  is:

$$\hat{F}[S^3] = \hat{F} \left[ \text{+1} \bigcirc \right] = \alpha^{-1} \sum_{\ell=0,1/2,1,\dots,k/2} \mu_\ell q^{C_\ell} \frac{S_{\ell 0}}{S_{00}}$$

where  $\mu_\ell = S_{0\ell}$  and the factor  $q^{C_\ell}$  is the effect from the framing +1 (one right-handed writhe). We make use of an identity:  $\sum_\ell S_{j\ell} q^{C_\ell} S_{\ell m} = \alpha q^{-C_j-C_m} S_{jm}$  which is closely related to the modular transformations of a torus. Thus this invariant for  $S^3$  is simply:

$$\hat{F}[S^3] = 1$$

For the three-manifold  $S^2 \times S^1$  (with surgery representation as an unknot with zero framing) is:

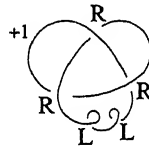
$$\hat{F}[S^2 \times S^1] = \hat{F} \left[ \begin{array}{c} 0 \\ \bigcirc \end{array} \right] = \sum_{\ell} \mu_{\ell} \frac{S_{\ell 0}}{S_{00}} = \sum_{\ell} \frac{S_{0\ell} S_{\ell 0}}{S_{00}} = \frac{1}{S_{00}}$$

where orthogonality property of the  $S$  matrix,  $\sum_{\ell} S_{j\ell} S_{\ell m} = \delta_{jm}$ , has been used.

Next for the three-dimensional real projective space  $RP^3$  (this is an  $S^3$  with antipodal points identified), the invariant is:

$$\hat{F}[RP^3] = \hat{F} \left[ \begin{array}{c} +2 \\ \text{trefoil} \end{array} \right] = \alpha^{-1} \sum_{j=0, \frac{1}{2}, 1, \dots, \frac{k}{2}} \frac{S_{0j} q^{2C_j} S_{j0}}{S_{00}}.$$

A slightly more complex example we take up is the Poincare manifold  $P^3$  (also known as dodecahedral space or Dehn's homology sphere). It is a homology three-sphere given by the set of points  $(u, v, w)$  in complex 3-space such that  $u^2 + v^3 + w^5 = 0$  and  $|u|^2 + |v|^2 + |w|^2 = 1$ . Its surgery presentation is given [21] by a right-handed trefoil knot with framing +1:



Notice, each right-handed crossing of the trefoil introduces +1 linking number between the knot and its vertical framing, and each of the two left-handed writhes contributes  $-1$  so that the total frame number of this knot is +1. Now using the knot invariant for trefoil in vertical framing given above, the invariant for this three-manifold can easily be written down:

$$\hat{F}[P^3] = \alpha^{-1} \sum_{j=0, \frac{1}{2}, 1, \dots, \frac{k}{2}} S_{0j} q^{-2C_j} \sum_{m=0, 1, \dots, \min(2\ell, k-2\ell)} (-)^m [2m+1] q^{-3C_j + \frac{3C_m}{2}}$$

The two left-handed writhes introduce a factor of  $q^{-2C_j}$ .

The invariant  $\hat{F}$  for a manifold  $M^3$  constructed above is same, up to a normalization, as the partition function of an  $SU(2)$  Chern-Simons theory on that manifold[5, 11]:

$$Z[M^3] = \hat{F}[M^3] S_{00}. \quad (16)$$

Generally, it is rather difficult to obtain the Chern-Simons partition function for a given three-manifold  $M^3$  directly. But, the formulae above, make its computation through  $\hat{F}$  rather easy.

The three-manifold invariant presented here is given in terms of link invariants from  $SU(2)$  Chern-Simons theory. It is clear that a similar construction can be done with link invariants from Chern-Simons gauge theories based on other semi-simple groups. This would yield a new method of obtaining the partition function of such Chern-Simons theories.

Next question we may ask is: Is this three-manifold invariant complete? Two manifolds  $M$  and  $M'$  for which the invariants  $\hat{F}[M]$  and  $\hat{F}[M']$  are different can not be homeomorphic to each other. But the converse is not always true; for two arbitrary manifold, the invariants need not be always different. Recall the invariants obtained from Chern-Simons theory for mutant knots are not distinct. Hence, manifold obtained by surgery on topologically inequivalent mutant knots can not be distinguished by this three-manifold invariant.

## 6 Perturbative non-Abelian Chern-Simons theory

Though Chern-Simons theories have been solved exactly and non-perturbatively as discussed above, perturbative analysis of these theories are also rich in their mathematical structure. If we expand the expectation value of the Wilson loop operator associated with a knot as a perturbative power

series in the coupling constant, the coefficients of such an expansion have a deep mathematical meaning. These on their own are topological invariants characteristic of the knot.

Last decade has also witnessed enormous research activity in direct perturbative calculations in Chern-Simons gauge field theory [26]. By simple power counting this theory is superrenormalizable. There are divergences, which need to be regularized. The effective coupling constant  $k$  does in general depend on the regularization. In a class of regularizations, a shift in the coupling constant takes place:  $k \rightarrow k + 2$  for  $SU(2)$  theory. This shift is consistent with the effective coupling in the non-perturbative studies of the theory.

It is very easy to see that the first order contribution to the vacuum expectation value of the Wilson loop operator for a knot is the self-linking number of the knot up to some group theoretic factors. This is so because at this order, the theory reduces to essentially Abelian Chern-Simons theories. Topological regularization of the coincident loop integrals through framing as discussed in Sec.3 earlier, leads to this result. Higher order contributions to the expectation value of a Wilson loop operator in an  $SU(2)$  Chern-Simons theory yield the famous *Vassiliev invariants*. These were first introduced by V.A. Vassiliev in 1990 from a totally different mathematical framework involving a study of the space of all smooth maps of  $S^1$  into  $S^3$ . These maps have different types of singularities. According to the type of singularities, this space of the maps divides into classes, each of which corresponds to a knot type. These classes are characterized by the families of invariants characterizing the knot [27].

Perturbative studies of Chern-Simons theory have provided new insights into the theory of Vassiliev invariants. In a gauge theory, perturbative calculations are to be performed in a definite gauge. Calculations in the Landau gauge [28] lead to covariant integral representations of Vassiliev invariants, also known as configuration space integrals first developed by Bott and Taubes in 1994 [29]. Another integral representation of the Vassiliev invariants was introduced by M. Kontsevich in 1993 [30]. This corresponds to perturbative calculation of the Chern-Simons theory in light-cone gauge [31]. It is rather very difficult to realize that these two integrals represent the same invariant. However, from a field theoretic point of view, this is simply a consequence of gauge invariance. Calculations in the temporal gauge have yielded yet another formulation of these invariants, leading to combinatorial formulae for them [32].

## 7 Gravity and Chern-Simons theory

While Chern-Simons theories have provided a powerful framework for theory of knots, these field theories are also of direct relevance in physics. For example there is an intimate relationship between these field theories and three dimensional gravity which is also a topological field theory. In fact two copies of  $SU(2)$  Chern-Simons theories represent gravity in Euclidean three-space with a negative cosmological constant [33]. To see this, just consider the partition function of two  $SU(2)$  Chern-Simons theories recast in terms of an  $SL(2, C)$  Chern-Simons theory as:

$$Z = \int [dA, d\bar{A}] \exp \left\{ \frac{ik}{8\pi} \int_{M^3} d^3x \epsilon^{\mu\nu\alpha} \left[ \text{tr} (A_\mu \partial_\nu A_\alpha + \frac{2}{3} A_\mu A_\nu A_\alpha) - \text{tr} (\bar{A}_\mu \partial_\nu \bar{A}_\alpha + \frac{2}{3} \bar{A}_\mu \bar{A}_\nu \bar{A}_\alpha) \right] \right\}$$

where  $A$  is an the  $SL(2, C)$  gauge field and  $\bar{A}$  its conjugate. This partition function is square of two  $SU(2)$  partition functions:  $Z_{SL(2,C)} = |Z_{SU(2)}|^2$ . Make a change of variables  $A = \omega + ie/\ell$  and  $\bar{A} = \omega - ie/\ell$ , where  $\omega$  and  $e$  are the gravitational spin connection and triad respectively. Writing  $kS[A] = \frac{k}{8\pi} \int d^3x \epsilon^{\mu\nu\alpha} \text{tr} [A_\mu \partial_\nu A_\alpha + \frac{2}{3} A_\mu A_\nu A_\alpha]$ , this then relates the action of these two Chern-Simons theories to Einstein-Hilbert action for three dimensional gravity:

$$ik(S[A] - S[\bar{A}]) = \frac{1}{16\pi G} \int_{M^3} d^3x \sqrt{g} \left( R + \frac{2}{\ell^2} \right) \quad (17)$$

where the cosmological constant  $= -1/\ell^2$  is negative and the Chern-Simons coupling is related to the gravitational coupling as  $k = \ell/(4G)$ .

This is closely related to another development in gravity. Three-dimensional gravity has a lattice formulation, first introduced by G. Ponzano and T. Regge in 1968 [34]. Here the three-manifold is decomposed into simplices. Each three-simplex is a tetrahedron. To each edge of the tetrahedron, a half-integral spin  $j$ , called its *colour*, is assigned so that its length is given by  $\sqrt{j(j+1)}$ . The spins on the three edges of each triangular face satisfy the triangular angular momentum inequality relations. The gravitational partition function is constructed in terms of a Racah-Wigner six- $j$  symbols for each tetrahedron in the simplicial decomposition of the manifold. For large spins, the six- $j$  symbols reproduce the ordinary gravitational action. Ponzano-Regge partition function suffers from a problem: it diverges as all possible spin values are allowed to live on the edges. This, therefore requires a regularization. A slightly more complex generalization of this lattice gravity model, which also provides this regularization, is related to a model first introduced by V.G. Turaev and O.Y. Viro [35]. It replaces the ordinary 6- $j$  symbols by their  $q$ -deformed analogues (with  $q$  as a root of unity). For large spin values, the  $q$ -six- $j$  symbol can be shown to give Regge action for a tetrahedron and represents Euclidean gravity action with a negative cosmological constant. The Turaev-Viro model would then be a quantum description of this three dimensional gravity.

For a triangulation of the three-manifold in terms of tetrahedra labeled by  $t$  and colouring  $j_e$  of its edges labeled by  $e$ , Turaev-Viro partition function for a manifold without boundary is given by the formula:

$$Z_{TV} = \sum_{\text{colourings } j_e \leq k/2} \prod_{\text{vertices}} \frac{1}{\Lambda} \prod_{\text{edges } e} (-1)^{2j_e} [2j_e + 1] \times \prod_{\text{tetrahedra } t} \exp\left(-i\pi \sum_i j_i(t)\right) \left\{ \begin{matrix} j_1(t) & j_2(t) & j_3(t) \\ j_4(t) & j_5(t) & j_6(t) \end{matrix} \right\}_q \quad (18)$$

The square bracket indicates the  $q$ -numbers, and curly brackets represent the  $q$ -6 $j$  symbol. The deformation parameter  $q$  is related to the Chern-Simons coupling by  $q = \exp[2\pi i/(k+2)]$  and  $\Lambda = -2(k+2)/(q^{1/2} - q^{-1/2})^2 = (S_{00})^{-2}$ . This partition function is naturally regularized and finite due the restriction on the spins living on the edges ( $j_e \leq k/2$ ) introduced by the fact that the deformation parameter is a root of unity. Further this partition function can be shown to be exactly square of an  $SU(2)$  Chern-Simons partition function,  $Z_{TV} = |Z_{SU(2)}|^2$ . This provides yet another representation for the Chern-Simons partition function.

Notice that the integration measure in the partition function of two Chern-Simons theories above is  $[dA, d\bar{A}]$ , whereas for the gravity partition function, it is  $[de, d\omega]$ . Since  $A = \omega + ie/\ell$  and  $\bar{A} = \omega - ie/\ell$ , the relation between the two involves  $1/\ell$  factors as the Jacobian. In fact in more exact treatment, it becomes clear that the Jacobian for this change of variables introduces exactly a factor of  $\Lambda$  for every vertex of the triangulation, so that the gravity partition function is just the Turaev-Viro partition function without the  $1/\Lambda$  factors:

$$Z_{grav} = \sum_{\text{colourings } j_e \leq k/2} \prod_{\text{edges } e} (-1)^{2j_e} [2j_e + 1] \times \prod_{\text{tetrahedra } t} \exp\left(-i\pi \sum_i j_i(t)\right) \left\{ \begin{matrix} j_1(t) & j_2(t) & j_3(t) \\ j_4(t) & j_5(t) & j_6(t) \end{matrix} \right\}_q \quad (19)$$

For a manifold with boundary, this expression has additional factors of  $\exp(i\pi j_b) \sqrt{[2j_b + 1]}$  for every boundary edge with a spin  $j_b$ . This partition function then is a functional of the boundary triangulation and spins of edges on the boundary.

There are many interesting questions which can be addressed in this framework for three-dimensional gravity. Some of these are: how does a black hole look in this formulation? What is its entropy? Analysis shows that a black hole (Banados-Teitelboim-Zanelli black hole) is given by a solid torus. Its horizon is given by the longitudinal circle at the core of this solid torus. The possible states associated with this black hole are the states associated with different triangulations of the black hole manifold, with the restriction that the longitudes have same circumference. It can

be shown that correct semi-classical behaviour of entropy is reproduced by states corresponding to all possible triangulations of such an Euclidean black hole [36]. The dominant contribution comes from the states at the horizon.

Chern-Simons theories have also played an important role in non-perturbative formulation of canonical quantum gravity in four dimensions [37]. In this approach, the physical states are given by spin-networks with associated graphs in three-space, where edges are labeled by  $SU(2)$  spins (colours) and vertices are given by interwinning operators. Quantum mechanical operators corresponding to lengths, areas and volumes all have discrete spectrum. It can be argued that the boundary degrees of freedom of a black hole, say Schwarzschild black hole, in four dimensional gravity can be described by a Chern-Simons theory[38, 39]. The action embodying the appropriate boundary conditions on the black hole horizon consists of, in addition to the Einstein-Hilbert action (in suitable variables), an  $SU(2)$  Chern-Simons gauge theory living on a coordinate chart of a constant finite cross-sectional area on the horizon. The Chern-Simons coupling  $k$  is proportional to this constant cross-sectional area. As the fundamental quantum excitations are polymer like, the horizon area is generated by the punctures where these spin-polymers pierce it. A bulk polymer state that gives the horizon its area in this manner has to be compatible with the surface states on the horizon itself. These boundary states are described by a quantum  $SU(2)$  Chern-Simons theory on the horizon. That is, the space of these boundary degrees of freedom is given by the space of states of Chern-Simons theory on a three-manifold with an  $S^2$  boundary with finitely many punctures on which spins live. The entropy of the black hole emerges from these boundary states. For large areas, where essentially  $U(1)$  subgroup of  $SU(2)$  contributes, the entropy is calculated by counting these states. Their number grows exponentially with horizon area yielding the semi-classical Bekenstein-Hawking expression for black hole entropy[39]. For finite areas, full  $SU(2)$  counting has to be done. This has been done by exploiting the relation between the boundary states of the Chern-Simons theory and the space of conformal blocks of the associated Wess-Zumino conformal field theory on the boundary 2-sphere, a relationship which played a crucial role in obtaining the link invariants in Sec.4. This yields an *exact* formula for entropy of a non-rotating black hole which for large areas reproduces the semi-classical formula, but for finite areas goes beyond the Bekenstein-Hawking result[40].

## 8 Summary and Concluding remarks

We have attempted here to indicate how quantum field theories, which have been successfully used to describe physics of fundamental interactions of Nature, can also be used to study geometry and topology of low dimensional manifolds. These developments not only provide new insights into old problems of topology of these manifolds but also have been responsible for profoundly interesting new mathematical results. These developments have make use of many of the recent developments in quantum field theories. The interaction between quantum physics and mathematics has enriched both.

Chern-Simons gauge field theory, a topological quantum field theory, provides a powerful framework for modern theory of knots and links in any three-manifold. This is one of the rare quantum field theories which can be explicitly and non-perturbatively solved. While Abelian Chern-Simons theory provides a simple description of linking and self-linking numbers of a link, non-Abelian theories are even richer. For every representation of any non-Abelian gauge group, there is a new link invariant. Jones polynomial associated with spin  $1/2$  representation in an  $SU(2)$  Chern-Simons theory, is the simplest example of such link invariants. Even more general invariants (*coloured invariants*) are obtained if we place different representations on the component knots. The framework is rich enough to discuss the knots and links not only in simple manifold like  $R^3$  or  $S^3$ , but any arbitrary three-manifold. Chern-Simons partition function is a particularly interesting three-manifold invariant for which a simple and efficient computational method is available now. Perturbative studies of Chern-Simons theory have given a new framework for describing Vassiliev invariants.

In the process of developing this framework for knot theory, new representations of braids also



have been obtained. The close connection that braids have with Yang-Baxter equation, has provided methods of obtaining a variety of new exactly solvable two-dimensional statistical mechanical models in physics[5]. These models are the higher vertex generalization of the six-vertex model of Lieb and Wu and 19-vertex model of Zamolodchikov and Fateev.

Chern-Simons field theories are also of direct interest in other areas of physics. One area where these have found profound application is quantum gravity. Three-dimensional gravity with a negative cosmological constant, itself a topological field theory, is essentially described by two  $SU(2)$  Chern-Simons theories. Micro-states of a black hole in the four dimensional spin-polymer gravity can also be modeled by a Chern-Simons theory. This allows an exact computation of black hole entropy going beyond the semi-classical result. These calculations so far have been done for non-rotating black holes only. These need to be extended for charged and rotating black holes, which requires certain amount of technical work. Further, while an exact formula for quantum entropy of a non-rotating black hole has been derived, a similar exact formula for the expectation value of the area operator in the Chern-Simons approach is not known. Also, a satisfactory understanding of Hawking radiation in this picture is yet to be developed.

String theory is another interesting framework in which black hole entropy has been analyzed in recent times. Though it provides a fundamental quantum description, unfortunately, calculations in this theory can be done for extremal or near extremal black holes only. These despite their mathematical interest are not astrophysically realistic. In particular, black holes of interest such as a Schwarzschild black hole are not generally amenable to analysis in this approach. Also supersymmetry plays an important role in the string picture. In contrast, modeling of micro-states of a black hole by an effective Chern-Simons theory is not limited by the constraint of extremality or near extremality. This framework handles the curved geometry of the black hole directly without invoking supersymmetry.

There are other topological quantum field theories also. One particularly interesting class is so called *cohomological field theories*. These are the field theoretical interpretations of four-manifold invariants obtained by S. Donaldson in 1983. His work is an example of developments in mathematics which have made critical use of some of the notions of physics [41]. His theory provides an understanding of the geometry in four dimensions through self-dual and anti-self-dual Yang-Mills gauge fields known to physicists as 'instantons and anti-instantons'. Five years later, E. Witten provided a quantum field theoretical framework for Donaldson's work in terms of a four dimensional topological Yang-Mills gauge field theory[42]. This field theory has certain kind of twisted supersymmetry. Donaldson invariants are given as the correlation functions in this field theory. In recent years, this area has registered even further boost through the work of Seiberg and Witten [43]. These developments use the powerful electric-magnetic duality to relate the cohomological field theory based on gauge group  $SU(2)$  to that based on  $U(1)$ . This brings in completely new insights into this area and makes calculation of Donaldson four manifold invariants rather easy.

## References

- [1] M. Atiyah: *The Geometry and Physics of Knots*, Cambridge Univ. Press (1989).
- [2] E. Witten: Commun. Math. Phys. **121** (1989) 351-399.
- [3] A.S. Schwarz: New topological invariants in the theory of quantized fields, Baku International Conference (1987).
- [4] R.K. Kaul: Complete solution of  $SU(2)$  Chern-Simons theory, hep-th/9212129; and Commun. Math. Phys. **162** (1994) 289 (hep-th/930532).
- [5] R.K. Kaul, Chern-Simons theory, knot invariants, vertex models and three-manifold invariants, hep-th/9804122, in *Frontiers of Field Theory, Quantum Gravity and Strings (Volume 227 in Horizons in World Physics)*, eds. R.K. Kaul et al, NOVA Science Publishers, New York (1999).

- [6] V.F.R. Jones: Bull. AMS (1985) 103-112; Ann. Math. **126** (1987) 335-388.
- [7] M. Wadati, T. Deguchi and Y. Akutsu: Phys. Rep. **180** (1989) 247 and references therein;  
M. Jimbo: Commun. Math. Phys. **102** (1986) 537;  
V.G. Turaev: Inv. Math. **92** (1988) 527.
- [8] A.N. Kirillov and N.Yu. Reshetikhin: Representation algebra  $U_q(SL(2))$ ,  $q$ -orthogonal polynomials and invariants of links, in *New Developments in the Theory of Knots*, ed. T. Kohno, World Scientific, Singapore (1989) .  
L. Alvarez Gaume, G. Gomez and G. Sierra: Phys. Letts. **B220** (1989) 142-152.  
C. Kassel, M. Rosso and V. Turaev: Quantum groups and knot invariants, Panoramas et syntheses 5, Societe Mathematique de France (1997).
- [9] N.Y. Reshtekhin and V. Turaev: Invent. Math. **103** (1991) 547.  
R. Kirby and P. Melvin: Invent. Math. **105** (1991) 473.
- [10] W.B.R. Lickorish: Math. Ann. **290** (1991) 657; and Pac. J. Math. **149** (1991) 337.
- [11] P. Ramadevi and Swatee Naik, Computation of Lickorish's three-manifold invariants using Chern-Simons theory, hep-th/9901061.
- [12] C. N. Little: Trans. R. Soc. Eddinburgh **39** (1990) 771;  
K.A. Perko: Proc. Am. Math. Soc. **45** (1974) 262.
- [13] J.W. Alexander: Trans. Am. Math. Soc. **30** (1928) 275.
- [14] P. Freyd, D. Yetter, J. Hoste, W.B.R. Lickorish, K. Millet and A. Ocneanu: Bull. AMS. **12** (1985) 239;  
J.H. Przytycki and K.P. Traczyk: Kobe J. Math. **4** (1987) 115.
- [15] William Thomas Kelvin, First Baron, *Mathematical and Physical Papers*, vol. IV, *Hydrodynamics and General Dynamics*, Cambridge University Press, Cambridge (1910).
- [16] R.K. Kaul and R. Rajaraman: Phys. Letts. **B249** (1990) 433-437.
- [17] G. Calugareanu: Rev. Math. Press. App. **4** (1959) 5 and Czech. Math. Jour. **11** (1961) 5881;  
W.F. Pohl: J. Math. Mech. **17** (1968) 975.
- [18] F.H. C. Crick: Proc. Natl. Acad. Sciences, USA, **73** (1971) 2639;  
F. Kamenetskii and A.V. Vologodskii: Sov. Phys. Usp. **24** (8) (1981) 679.
- [19] J.W. Alexander: Proc. Natl. Acad. **9** (1923) 93.
- [20] J.S. Birman: *Braids, Links and Mapping Class groups*, Annals of Mathematics Studies Princeton Univ. Press (1975).
- [21] D. Rolfsen: *Knots and links*, Publish or Perish, Berkeley (1976).
- [22] P. Ramadevi, T.R. Govindarajan and R.K. Kaul: Mod. Phys. Letts. **A9** (1994) 3205.
- [23] P. Ramadevi, T.R. Govindarajan and R.K. Kaul: Mod. Phys. Letts. **A10** (1995) 1635.
- [24] A.D. Wallace: Can. J. Math. **12** (1960) 503; B.R. Lickorish: Ann. of Math. **76** (1962) 531.
- [25] R. Kirby: Invent. Math. **45** (1978) 35;  
R. Fenn and C. Rourke: **18** (1979) 1.
- [26] For a recent review see, J.M.F. Labastida: Chern-Simons gauge theory: ten years after, hep-th/9905057.
- [27] V.A. Vassiliev: Cohomology of knot spaces, *Theory of singularities and its applications, Advances in Soviet Mathematics*, vol. 1, American Math. Soc., Providence, RI (1990) 23-69.

- [28] E. Guadagni, M. Martellini and M. Mintchev: Phys. Letts **B227** (1989) 111 and Nucl. Phys. **B330** (1990) 575;  
M. Alvarez and J.M.F. Labastida: Nucl. Phys. **395** (1993) 198 and Nucl. Phys. **B433** (1995) 555;  
D. Althschuler and L. Friedel: Commun. Math. Phys. **187** (1997) 261 and **170** (1995) 41.
- [29] R. Bott and C. Taubes: Jour. Math. Phys. **35** (1994) 5247.
- [30] M Kontsevich: Advances in Soviet Math. **16**, Part 2 (1993) 137.
- [31] A.S. Cattaneo, P. Cotta-Ramusino, J. Frohlich and M. Martellini: J. Math. Phys. **36** (1995) 6137;  
J.M.F. Labastida and E. Perez: J. Math. Phys. **39** (1998) 5183.
- [32] J.M.F. Labastida and E. Perez: Combinatorial formulae for Vassiliev invariants from Chern-Simons gauge theory, hep-th/9807155 and Vassiliev invariants in the context of Chern-Simons gauge theory, hep-th/9812105.
- [33] E. Witten: Nucl. Phys. **B311** (1989) 46.
- [34] G. Ponzano and T. Regge in *Spectroscopic and group theoretical methods in physics*, ed. F. Block, North-Holland, Amsterdam 1968.
- [35] V.G. Turaev and O.Y. Viro: Topology **31** (1992) 865;  
H. Ooguri: Nucl. Phys. **B382** (1992) 865.
- [36] V. Suneeta, R.K. Kaul and T.R. Govindarajan: gr-qc/9811071, Mod. Phys. Letts. **A14** (1999) 349.
- [37] A. Ashtekar: Phys. Rev. Lett. **57** (1986) 2244;  
C. Rovelli and L. Smolin: Nucl. Phys. **B331** (1990) 80;  
A. Ashtekar and J. Lewandowski: in *Knots and quantum gravity*, ed. J. Baez, Oxford Univ. Press, 1994;  
J. Baez: Lett. Math. Phys. **31** (1994) 213;  
C. Rovelli and L. Smolin: Phys. Rev. **D52** (1995) 5743.
- [38] L. Smolin: J. Math. Phys. **36** (1995) 6417;  
A. Balachandran, L. Chandar and A. Momen: Nucl. Phys. **B461** (1996) 581;  
C. Rovelli: Phys. Rev. Lett. **77** (1996) 3288.
- [39] A. Ashtekar, J. Baez, A. Corichi, K. Krasnov: Phys. Rev. Lett. **80** (1998) 904.
- [40] R.K. Kaul and P. Majumdar: Phys. Letts. **B439** (1998) 267;  
P. Majumdar: Indian Jour. of Phys. **73B** (1999) 147.
- [41] S.K. Donaldson: J. Diff. Geom. **18** (1983) 269 and Polynomial invariants for smooth four-manifolds, Topology **29** (1990) 257.
- [42] E. Witten: Commun. Math. Phys. **117** (1988) 353.
- [43] N. Seiberg and E. Witten: Nucl. Phys. **B426** (1994) 19.

This article has been reproduced from  
*Commun. Math. Phys.* 121,351-399 (1989)  
with kind permission from Springer-Verlag

## Quantum Field Theory and the Jones Polynomial \*

Edward Witten \*\*

School of Natural Sciences, Institute for Advanced Study, Olden Lane, Princeton,  
NJ 08540, USA

**Abstract.** It is shown that  $2 + 1$  dimensional quantum Yang-Mills theory, with an action consisting purely of the Chern-Simons term, is exactly soluble and gives a natural framework for understanding the Jones polynomial of knot theory in three dimensional terms. In this version, the Jones polynomial can be generalized from  $S^3$  to arbitrary three manifolds, giving invariants of three manifolds that are computable from a surgery presentation. These results shed a surprising new light on conformal field theory in  $1 + 1$  dimensions.

In a lecture at the Hermann Weyl Symposium last year [1], Michael Atiyah proposed two problems for quantum field theorists. The first problem was to give a physical interpretation to Donaldson theory. The second problem was to find an intrinsically three dimensional definition of the Jones polynomial of knot theory. These two problems might roughly be described as follows.

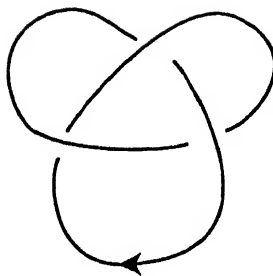
Donaldson theory is a key to understanding geometry in four dimensions. Four is the physical dimension at least macroscopically, so one may take a slight liberty and say that Donaldson theory is a key to understanding the geometry of space-time. Geometers have long known that (via de Rham theory) the self-dual and anti-self-dual Maxwell equations are related to natural topological invariants of a four manifold, namely the second homology group and its intersection form. For a simply connected four manifold, these are essentially the only classical invariants, but they leave many basic questions out of reach. Donaldson's great insight [2] was to realize that moduli spaces of solutions of the self-dual Yang-Mills equations can be powerful tools for addressing these questions.

Donaldson theory has always been an intrinsically four dimensional theory, and it has always been clear that it was connected with mathematical physics at least at the level of classical nonlinear equations. The puzzle about Donaldson theory was whether this theory was tied to more central ideas in physics, whether it could be interpreted in terms of quantum field theory. The most important

---

\* An expanded version of a lecture at the IAMP Congress, Swansea, July, 1988

\*\* Research supported in part by NSF Grant No. 86-20266, and NSF Waterman Grant 88-17521



**Fig. 1.** A knot in three dimensional space

evidence for the existence of such a connection had to do with Floer's work on three manifolds [3] and the nature of the relation between Donaldson theory and Floer theory. Also, the "Donaldson polynomials" had an interesting formal analogy with quantum field theory correlation functions. It has turned out that Donaldson theory can indeed be given a physical interpretation [4].

As for the Jones polynomial and its generalizations [5–11], these deal with the mysteries of knots in three dimensional space (Fig. 1). The puzzle on the mathematical side was that these objects are invariants of a three dimensional situation, but one did not have an intrinsically three dimensional definition. There were many elegant definitions of the knot polynomials, but they all involved looking in some way at a two dimensional projection or slicing of the knot, giving a two dimensional algorithm for computation, and proving that the result is independent of the chosen projection. This is analogous to studying a physical theory that is in fact relativistic but in which one does not know of a manifestly relativistic formulation – like quantum electrodynamics in the 1930's.

On the physical side, the puzzle about the knot polynomials was the following. Unlike the Donaldson theory, where a connection with quantum field theory was not obvious, the knot polynomials have been intimately connected almost from the beginning with two dimensional many body physics. In fact, constructions of the knot polynomials have related them to two dimensional (or  $1 + 1$  dimensional) many-body physics in a bewildering variety of ways, mainly involving soluble lattice models [7], solutions of the Yang-Baxter equation [8], and monodromies of conformal field theory [11]. In the latter interpretation, the knot polynomials are related to aspects of conformal field theory that have been particularly fruitful recently [12–16]. On the statistical mechanical side, studies of the knot polynomials have related them to Temperley-Lieb algebras and their generalizations, and to other aspects of soluble statistical mechanics models in  $1 + 1$  dimensions. For physicists the challenge of the knot polynomials has been to bring order to this diversity, find the unifying themes, and learn what it is that is three dimensional about two dimensional conformal field theory.

Now, the Donaldson and Jones (and Floer and Gromov [17]) theories deal with topological invariants, and understanding these theories as quantum field theories involves constructing theories in which all of the observables are topological invariants. Some physicists might consider this to be a little bit strange, so let us pause to explain the physical meaning of "topological invariance". The physical meaning is really "general covariance". Something that can be computed from a manifold  $M$  as a topological space (perhaps with a

smooth structure) without a choice of metric is called a “topological invariant” (or a “smooth invariant”) by mathematicians. To a physicist, a quantum field theory defined on a manifold  $M$  without any a priori choice of a metric on  $M$  is said to be generally covariant. Obviously, any quantity computed in a generally covariant quantum field theory will be a topological invariant. Conversely, a quantum field theory in which all observables are topological invariants can naturally be seen as a generally covariant quantum field theory. Indeed, the Donaldson, Floer, Jones, and Gromov theories can be seen as generally covariant quantum field theories in four, three, and two space-time dimensions. The surprise, for physicists, perhaps comes in how general covariance is achieved. General relativity gives us a prototype for how to construct a quantum field theory with no a priori choice of metric – we introduce a metric, and then integrate over all metrics. This example is so influential in our thinking that we tend to think of a generally covariant theory as being, by definition, a theory in which the metric is a dynamical variable. The lesson from the Donaldson, Floer, Jones, and Gromov theories is precisely that there are highly non-trivial quantum field theories in which general covariance is realized in other ways. In particular, in this paper we will describe an exactly soluble generally covariant quantum field theory in which general covariance is achieved not by integrating over metrics but because we begin with a gauge invariant Lagrangian that does not contain a metric.

## 1. The Chern-Simons Action

We have been urged [1] to try to interpret the Jones polynomial in terms of three dimensional Yang-Mills theory. So we begin on an oriented three manifold  $M$  with a compact simple gauge group  $G$ . We pick a  $G$  bundle  $E$ , which may as well be trivial, and on  $E$  we place a connection  $A_i^a$ , which can be viewed as a Lie algebra valued one form ( $a$  runs over a basis of the Lie algebra, and  $i$  is tangent to  $M$ ). An infinitesimal gauge transformation is

$$A_i \rightarrow A_i - D_i \varepsilon, \quad (1.1)$$

where  $\varepsilon$ , a generator of the gauge group, is a Lie algebra valued zero form and the covariant derivative is  $D_i \varepsilon = \partial_i \varepsilon + [A_i, \varepsilon]$ . The curvature is the Lie algebra valued two form  $F_{ij} = [D_i, D_j] = \partial_i A_j - \partial_j A_i + [A_i, A_j]$ . Now we need to choose a Lagrangian. We will *not* pick the standard Yang-Mills action<sup>1</sup>

$$\mathcal{L}_0 = \int_M \sqrt{g} g^{ik} g^{jl} \text{Tr}(F_{ij} F_{kl}), \quad (1.2)$$

as this depends on the choice of a metric  $g_{ij}$ . We want to formulate a generally covariant theory (in which all observables will be topological invariants), and to this aim we want to pick a Lagrangian which does not require any choice of metric.

<sup>1</sup> In what follows, the symbol “Tr” denotes an invariant bilinear form on the Lie algebra of  $G$ , a multiple of the Cartan-Killing form; we will specify the normalization presently

Precisely in three dimensions there is a reasonable choice, namely the integral of the Chern-Simons three form:

$$\begin{aligned}\mathcal{L} &= \frac{k}{4\pi} \int_M \text{Tr}(A \wedge dA + \frac{2}{3} A \wedge A \wedge A) \\ &= \frac{k}{8\pi} \int_M \varepsilon^{ijk} \text{Tr}(A_i(\partial_j A_k - \partial_k A_j) + \frac{2}{3} A_i[A_j, A_k]).\end{aligned}\quad (1.3)$$

The Chern-Simons term in three dimensional gauge theory has a relatively long history. The abelian gauge theory with only a Chern-Simons term was studied by Schwarz [18] and in unpublished work by I. Singer. Three dimensional gauge theories with the Chern-Simons term added to the usual action (1.2) were introduced in [19–21]. The nonabelian theory with only Chern-Simons action was studied classically by Zuckerman [22]. The abelian Chern-Simons theory has recently been studied in relation to fractional statistics by Hagen [24] and by Arovas et al. [25] and in relation to linking numbers by Polyakov [23] and Fröhlich [15]. The novelty in our present discussion is that we will consider the quantum field theory defined by the nonabelian Chern-Simons action and argue that it is exactly soluble and has important implications for three dimensional geometry and two dimensional conformal field theory.

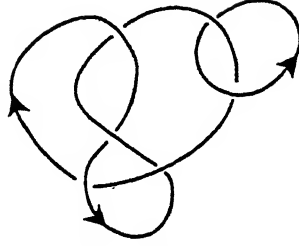
The first fundamental property of the Chern-Simons theory is the quantization law first discussed in [21]. It arises because the group  $\hat{G}$  of continuous maps  $M \rightarrow G$  is not connected. In the homotopy classification of such maps one meets at least the fact that  $\pi_3(G) \simeq \mathbb{Z}$  for every compact simple group  $G$ . Though (1.3) is invariant under the component of the gauge group that contains the identity, it is not invariant under gauge transformations of non-zero “winding number”, gauge transformations associated with non-zero elements of  $\pi_3(G)$ . Under a gauge transformation of winding number  $m$ , the transformation law of (1.3) is

$$\mathcal{L} \rightarrow \mathcal{L} + \text{const} \cdot m. \quad (1.4)$$

As in Dirac’s famous work on magnetic monopoles, consistency of quantum field theory does not quite require the single-valuedness of  $\mathcal{L}$ , but only of  $\exp(i\mathcal{L})$ . For this purpose, it is necessary and sufficient that the “constant” in (1.4) should be an integral multiple of  $2\pi$ . This gives a quantization condition on the parameter called  $k$  in (1.3). If  $G$  is  $SU(N)$  and “Tr” means a trace in the  $N$  dimensional representation, then the requirement is that  $k$  should be an integer. In general, for any  $G$ , we can uniquely fix the so far unspecified normalization of “Tr” so that the quantization condition is  $k \in \mathbb{Z}$ .

We will see later that  $k$  is very closely related to the central charge in the theory of highest weight representations of affine Lie algebras. It is no accident that the reasoning which shows that  $k$  must be quantized in (1.3) has a  $1+1$  dimensional analogue [26] which leads to quantization of the central charge in the representation theory of affine algebras.

In quantum field theory, in addition to a Lagrangian, one also wishes to pick a suitable class of gauge invariant observables. In the present context, the usual gauge invariant local operators would not be appropriate, as they spoil general covariance. However, the “Wilson lines” so familiar in QCD give a natural class of



**Fig. 2.** Several linked but non-intersecting oriented knots in a three manifold  $M$ . Such a collection of knots is called a “link”

gauge invariant observables that do not require a choice of metric. Let  $C$  be an oriented closed curve in  $M$ . Intrinsically  $C$  is simply a circle, but the topological classification of embeddings of a circle in  $M$  is very complicated, as we observe in Fig. 1. Let  $R$  be an irreducible representation of  $G$ . One then defines the “Wilson line”  $W_R(C)$  to be the following functional of the connection  $A_i$ . One computes the holonomy of  $A_i$  around  $C$ , getting an element of  $G$  that is well-defined up to conjugacy, and then one takes the trace of this element in the representation  $R$ . Thus, the definition is

$$W_R(C) = \text{Tr}_R P \exp \int_C A_i dx^i. \quad (1.5)$$

The crucial property of this definition is that there is no need to introduce a metric, so general covariance is maintained.

We now can formulate the general problem of interest. In an oriented three manifold  $M$ , we take  $r$  oriented and non-intersecting knots  $C_i$ ,  $i = 1 \dots r$ , whose union is what knot theorists would call a “link”  $L$ . We assign a representation  $R_i$  to each  $C_i$ , and we propose to calculate the Feynman path integral

$$\int D\mathcal{A} \exp(i\mathcal{L}) \prod_{i=1}^r W_{R_i}(C_i). \quad (1.6)$$

The symbol  $D\mathcal{A}$  represents Feynman’s integral over all gauge orbits, that is, an integral over all equivalence classes of connections modulo gauge transformations. Of course, (1.6) has exactly the formal structure of some familiar observables in QCD, the difference being that we are in three dimensions instead of four and we have chosen a somewhat exotic gauge theory action. We will call (1.6) the “partition function” of  $M$  with the given link, or the (unnormalized) “expectation value” of the given link; we will denote it as  $Z(M; C_i, R_i)$  or simply as  $Z(M; L)$  for short.

For the case of links in  $S^3$ , we will claim that the invariants (1.6) are exactly those that appear in the Jones theory and its generalizations. Simply replacing  $S^3$  with a general oriented three manifold  $M$  gives a very intriguing (and as we will see, effectively computable) generalization of the known knot polynomials. Taking  $r = 0$  (no knots), (1.6) gives invariants of the oriented three manifold  $M$  which also turn out to be effectively computable. Before getting into any details, let us note a few preliminary indications of a possible connection between (1.6) and the Jones theory:

(1) In (1.6) we see the right variables, namely a compact Lie group  $G$ , a choice of representation  $R_i$  for each component  $C_i$  of the link  $L$ , and an additional



variable  $k$ . [In knot theory one usually makes an analytic continuation and replaces  $k$  by a complex variable  $q$ , but it has been known since Jones' original work that there are special properties at special values of  $q$ . We claim that these properties reflect the fact that the three dimensional gauge theory with action (1.3) is well-defined only if  $k$  is an integer.] The two variable generalization of the Jones polynomial corresponds to the case that  $G$  is  $SU(N)$ , and the  $R_i$  are all the defining  $N$  dimensional representation of  $SU(N)$ . The two variables are  $N$  and  $k$ , analytically continued to complex values. The Kauffman polynomial similarly arises for  $G = SO(N)$  and  $R$  the  $N$  dimensional representation.

(2) As a further check on the plausibility of a relation between (1.6) and the knot polynomials, let us note first of all that (1.6) depends on a choice of the orientation of  $M$ , as this enters in fixing the sign of the Chern-Simons form. Likewise, (1.6) depends on the orientations of the  $C_i$ , since these enter in defining the Wilson lines (in computing the holonomy around  $C_i$ , one must decide in which direction to integrate around  $C_i$ ). If, however, one reverses the orientation of one of the  $C_i$  and simultaneously exchanges the representation  $R_i$  with its complex conjugate  $\bar{R}_i$ , then the definition of the Wilson lines is unchanged, so (1.6) is invariant under this process. And if (without changing the  $R_i$ ) one reverses the orientations of *all* components  $C_i$  of the link  $L$ , then (1.6) is unchanged because of a symmetry that physicists would call "charge conjugation". This is an involution of the Lie algebra of  $G$  that exchanges all representations with their complex conjugates; applying this involution to all integration variables in (1.6) leaves (1.6) invariant while exchanging all  $R_i$  with their conjugates or equivalently reversing the orientation of all the  $C_i$ . These are important formal properties of the knot polynomials.

## 2. The Weak Coupling Limit

To begin with, since a non-abelian gauge theory with only a Chern-Simons action may seem unfamiliar, one might ask whether this Lagrangian really does lead to a sensible quantum theory, and really can be regulated to give topologically invariant results. In this section, we will briefly investigate this point by studying the theory in a weak coupling limit in which computations are comparatively straightforward. This is the limit of large  $k$ .<sup>2</sup> For large  $k$ , the path integral

$$Z = \int D\mathcal{A} \exp \left( \frac{ik}{4\pi} \int_M \text{Tr} \left( A \wedge dA + \frac{2}{3} A \wedge A \wedge A \right) \right) \quad (2.1)$$

(for the moment we omit knots) contains an integrand which is wildly oscillatory. The large  $k$  limit of such an integral is given by a sum of contributions from the points of stationary phase. The stationary points of the Chern-Simons action are precisely the "flat connections", that is, the gauge fields for which the curvature vanishes

$$F_{ij}^a = 0. \quad (2.2)$$

---

<sup>2</sup> The reader may wish to bear in mind that the discussion in this section and the next contains a number of technicalities which are part of the logical story but perhaps not essential on a first reading

Gauge equivalence classes of such flat connections correspond to homomorphisms

$$\phi: \pi_1(M) \rightarrow G, \quad (2.3)$$

or more exactly to equivalence classes of such homomorphisms, up to conjugation. If for simplicity we suppose that the topology of  $M$  is such that there are only finitely many classes of homomorphisms (2.3), then the large  $k$  behavior of (2.1) will be a sum

$$Z = \sum_{\alpha} \mu(A^{(\alpha)}), \quad (2.4)$$

where the  $A^{(\alpha)}$  are a complete set of gauge equivalence classes of flat connections, and  $\mu(A^{(\alpha)})$  is to be obtained by stationary phase evaluation of (2.1), expanding around  $A^{(\alpha)}$ . This reduction to a stationary phase evaluation means that the nonabelian theory, for large  $k$ , is closely related to the abelian theory. This in turn has been shown [18] to lead to Ray-Singer analytic torsion [27], which is closely related to the purely topological Reidemeister torsion. The  $\mu(A^{(\alpha)})$  may be evaluated as follows. We make in (2.1) the change of variables  $A_i = A_i^{(\alpha)} + B_i$ , where  $B_i$  is the new integration variable. An important invariant of the flat connection  $A^{(\alpha)}$  is its Chern-Simons invariant

$$I(A^{(\alpha)}) = \frac{1}{4\pi} \int_M \text{Tr} (A^{(\alpha)} \wedge dA^{(\alpha)} + \frac{2}{3} A^{(\alpha)} \wedge A^{(\alpha)} \wedge A^{(\alpha)}). \quad (2.5)$$

When the Chern-Simons action is expanded in powers of  $B_i$ , the first terms are

$$\mathcal{L} = k \cdot I(A^{(\alpha)}) + \frac{k}{4\pi} \int_M \text{Tr} (B \wedge DB). \quad (2.6)$$

Here it is understood that in (2.6), the expression  $DB$  denotes the covariant exterior derivative of  $B$  with respect to the background gauge field  $A^{(\alpha)}$ ; it does not depend on a metric on  $M$ . A salient point is that in (2.6) there is no term linear in  $B$ , since  $A^{(\alpha)}$  is a critical point of the action.

To carry out the Gaussian integral in (2.6), gauge fixing is needed. There is no way to carry out this gauge fixing without picking a metric on  $M$  (or in some other way breaking the symmetry of the problem). After picking such a metric, a convenient gauge choice is  $D_i B^i = 0$  (with  $D_i$  the covariant derivative constructed from the metric and the background gauge field  $A^{(\alpha)}$ ). The standard Faddeev-Popov construction then gives rise to a gauge fixing Lagrangian

$$\mathcal{L}_{\text{gauge}} = \int_M (\text{Tr} \phi D_i B^i + \text{Tr} \bar{c} D_i D^i c). \quad (2.7)$$

Here  $\phi$  is a Lagrangian multiplier that enforces the gauge condition  $D_i B^i = 0$ , and  $c, \bar{c}$  are anticommuting “ghosts” that are introduced to get the right measure on the space of gauge fields modulo gauge transformations. The quadratic terms in  $\phi$  and  $B$  that can be found in (2.6), (2.7) have a natural geometric interpretation, described (in the abelian case) in [18]. Let  $D$  be the exterior derivative on  $M$ , twisted by the flat connection  $A^{(\alpha)}$ , and let  $*$  be the Hodge operator that maps  $k$  forms to  $3 - k$  forms. On a three manifold one has a natural self-adjoint operator  $L = *D + D*$  which maps differential forms of even order to forms of even order

and forms of odd order to forms of odd order. Let  $L_-$  denote its restriction to forms of odd order. With  $B$  and  $\phi$  regarded as a one form and a three form, respectively, the boson kinetic operator in (2.6), (2.7) is precisely this operator  $L_-$ . The kinetic operator of the ghosts is also a natural geometrical operator, the Laplacian, which we will call  $\Delta$ . We can now give a formula for the stationary point contributions  $\mu(A^{(\alpha)})$  that appear in (2.4). This is

$$\mu(A^{(\alpha)}) = \exp(ikI(A^{(\alpha)})) \cdot \frac{\det(\Delta)}{\sqrt{\det(L_-)}}. \quad (2.8)$$

The phase factor in (2.8) is the value of the integrand in (2.1) at the point of stationary phase, and the determinants (whose absolute values can be defined by zeta functions) result from the Gaussian integral over  $B, \phi, c$ , and  $\bar{c}$ .

Now we come to the crucial point. To regularize the path integral, we have had to pick a Riemannian metric on  $M$ . Therefore, it is not obvious a priori that the  $\mu(A^{(\alpha)})$  computed this way will really be topological invariants. Perhaps the Chern-Simons theory suffers from anomalies, and cannot be regularized in a generally covariant fashion. Happily, we can now appeal to [18], where it was shown (in the context of the abelian theory, but this aspect of [18] generalizes) that the absolute value of the ratio of determinants appearing in (2.8) is precisely the Ray-Singer analytic torsion of the flat connection  $A^{(\alpha)}$ , and so in particular is a topological invariant. (The phase of this ratio of determinants is more delicate, and will be discussed later.) This is the first indication that topological invariants really can be obtained from the Chern-Simons theory.

### *The Phase of the Determinant*

Though the absolute value of the ratio of determinants in (2.8) is the analytic torsion discussed long ago by Schwarz, the phase requires additional study. The ghost determinant  $\det \Delta$  is real and positive, so the real issue is to study the phase of  $\det L_-$ . Because the operator  $L_-$  can be interpreted as a twisted Dirac operator, the phase of its determinant can be related to the study of the phase of odd dimensional fermion determinants, as studied by various authors [28]. However, I will here give a brief derivation of the relevant facts from the bosonic point of view, which is perhaps more natural in the present context. After an irrelevant rescaling of  $B$  and  $\phi$ , the integral of interest is

$$\int DBD\phi \exp\left(i \int_M \text{Tr}(B \wedge DB + \phi D * B)\right). \quad (2.9)$$

Upon changing variables to an orthonormal basis of eigenfunctions  $x_i$  of the operator  $L_-$ , with eigenvalues  $\lambda_i$ , (2.9) becomes

$$\prod_i \int_{-\infty}^{\infty} \frac{dx_i}{\sqrt{\pi}} e^{i\lambda_i x_i^2}. \quad (2.10)$$

Therefore the crucial integral to understand is

$$I = \int_{-\infty}^{\infty} \frac{dx}{\sqrt{\pi}} e^{i\lambda x^2}, \quad (2.11)$$

for real  $\lambda$ . We consider this integral to be defined by taking the limit as  $\varepsilon \rightarrow 0$  of the absolutely convergent integral

$$\int_{-\infty}^{\infty} \frac{dx}{\sqrt{\pi}} e^{i\lambda x^2} : e^{-\varepsilon x^2}. \quad (2.12)$$

With this or any other physically reasonable definition, the integral (2.11) is

$$I = \frac{1}{|\sqrt{\lambda}|} \cdot \exp\left(\frac{i\pi}{4} \text{sign } \lambda\right). \quad (2.13)$$

The phase of the path integral is thus proportional to  $\sum_i \text{sign } \lambda_i$ , or better, to its regularized version which is the “eta invariant” of Atiyah et al. [29]:

$$\eta(A^{(\alpha)}) = \frac{1}{2} \lim_{s \rightarrow 0} \sum_i \text{sign } \lambda_i |\lambda_i|^{-s}. \quad (2.14)$$

Thus, the phase of the path integral may be expressed in the formula

$$\frac{1}{\sqrt{\det L_-}} = \frac{1}{|\sqrt{\det L_-}|} \cdot \exp\left(\frac{i\pi}{2} \eta(A^{(\alpha)})\right). \quad (2.15)$$

This can be made more explicit by using the Atiyah-Patodi-Singer theorem, which for our purposes can be regarded as a formula that expresses the dependence of  $\eta$  on the flat connection  $A^{(\alpha)}$  about which we are expanding. In fact, in the case of the operator  $L_-$ , the formula is

$$\frac{1}{2} (\eta(A^{(\alpha)}) - \eta(0)) = \frac{c_2(G)}{2\pi} \cdot I(A^{(\alpha)}). \quad (2.16)$$

Here  $I(A^{(\alpha)})$  is the Chern-Simons invariant of the flat connection  $A^{(\alpha)}$ , as defined in (2.5),  $\eta(0)$  is the eta invariant of the trivial gauge field  $A = 0$ , and  $c_2(G)$  is the value of the quadratic Casimir operator of the group  $G$  in the adjoint representation, normalized so that  $c_2(SU(N)) = 2N$ . The effect of this factor is to replace  $k$  in (2.8) by  $k + c_2(G)/2$ ; in fact, the partition function (2.4) may now be written

$$Z = e^{i\pi\eta(0)/2} \cdot \sum_{\alpha} e^{i(k + c_2(G)/2) I(A^{(\alpha)})} \cdot T_{\alpha} \quad (2.17)$$

with  $T_{\alpha}$  [the absolute value of the ratio of determinants in (2.8)] being the torsion invariant of the flat connection  $A^{(\alpha)}$ .

Unfortunately, although  $I(A^{(\alpha)})$  and  $T_{\alpha}$  are topological invariants,  $\eta(0)$  is not; it depends on the choice of a metric on  $M$  in gauge fixing. Thus, to make sense of the phase of (2.17) requires further discussion, in the next subsection.

Before launching into that technical discussion, let us note that the computation just sketched actually has a very interesting spin-off. The fact that  $k$  in (2.8) has been replaced by  $k + c_2(G)/2$  in (2.17) appears to be the beginning of an explanation of the fact that in many formulas of 1 + 1 dimensional current algebra, quantum corrections have the effect of replacing  $k$  by  $k + c_2(G)/2$ . In turn, this is probably related to the fact that in various integrable models in 1 + 1

dimensions, such as the sine-gordon model, the WKB approximation is exact if one makes suitable and seemingly *ad hoc* changes in the values of the parameters, analogous to replacing  $k$  by  $k + c_2(G)/2$ .

### Trivialization of the Tangent Bundle

Now, let us discuss how the mysterious phase factor  $e^{i\pi\eta(0)/2}$  in (2.17) should be interpreted.

First of all,  $\eta(0)$  is the  $\eta$  invariant of the  $L_-$  operator coupled to (i) some metric  $g$  on  $M$ , and (ii) the trivial gauge field  $A = 0$ . Let  $d = \dim G$  be the dimension of the gauge group  $G$ . Since the gauge field is trivial, the  $L_-$  operator consists of  $d$  copies of the purely gravitational  $L_-$  operator coupled to the metric only. Thus, as a preliminary, we write

$$\eta(0) = d \cdot \eta_{\text{grav}}, \quad (2.18)$$

where  $\eta_{\text{grav}}$  is the eta invariant of the purely gravitational operator. Our problematical phase factor is

$$A = \exp\left(\frac{id\pi}{2} \cdot \eta_{\text{grav}}\right). \quad (2.19)$$

Now, with a particular regularization of the Chern-Simons quantum field theory, we have obtained the formula (2.17) which contains the ambiguous phase factor  $A$ . The goal is to find a different regularization which will preserve general covariance. Two regularizations should differ by a local counterterm, and in this case, since the problem phase (2.19) depends on the background metric only, we want a counterterm that depends on the background metric only. It is easy to see that the counterterm with the right properties is a multiple of the gravitational Chern-Simons term, which is defined (by analogy with the Yang-Mills Chern-Simons term) as

$$I(g) = \frac{1}{4\pi} \int_M \text{Tr}(\omega \wedge d\omega + \frac{2}{3}\omega \wedge \omega \wedge \omega). \quad (2.20)$$

Here  $\omega$  is the Levi-Civita connection on the spin bundle of  $M$ .<sup>3</sup>  $I(g)$  suffers from an ambiguity just similar to that of the Yang-Mills Chern-Simons action. To define  $I(g)$  as a number, one requires a trivialization of the tangent bundle of  $M$ . Although the tangent bundle of a three manifold can be trivialized, there is no canonical way to do this. Any two trivializations differ by an invariantly defined integer, which is the number of relative “twists”. The gravitational Chern-Simons functional has the property that if the trivialization of the tangent bundle of  $M$  is twisted by  $s$  units,  $I(g)$  transforms by

$$I(g) \rightarrow I(g) + 2\pi s. \quad (2.21)$$

Now, the Atiyah-Patodi-Singer theorem says that the combination

$$\frac{1}{2} \eta_{\text{grav}} + \frac{1}{12} \cdot \frac{I(g)}{2\pi} \quad (2.22)$$

<sup>3</sup> (2.20) is not the integral of an intrinsic local functional, so it would not usually arise as a counterterm. Whether or not “counterterm” is the right word, we will have to view (2.20) as a correction that must be added to the action if one wishes to work in the gauge  $D_i A^i = 0$

is a topological invariant, depending that is on the oriented three manifold  $M$  with a choice of trivialization of the tangent bundle, but not on the metric of  $M$ .<sup>4</sup> It is clear, therefore, what we must do. We replace  $\eta(0)/2$  in (2.17) by  $d$  times the combination that appears in (2.22) [the factor of  $d$  is the one that entered in (2.19)], so (2.17) is replaced by

$$Z = \exp \left( i\pi d \left( \frac{\eta_{\text{grav}}}{2} + \frac{1}{12} \cdot \frac{I(g)}{2\pi} \right) \right) \cdot \sum_{\alpha} e^{i(k + c_2(G)/2) I(A^{(\alpha)})} \cdot T_{\alpha}. \quad (2.23)$$

So, finally, we can see that the Chern-Simons partition function, at least for large  $k$ , can be defined as a topological invariant of the oriented, framed three manifold  $M$  (a framed three manifold being one that is presented with a homotopy class of trivializations of the tangent bundle).

The fact that it is necessary to specify a framing of the three manifold may look like a nuisance, but there is no real loss of information. From (2.21) we see that if the framing is shifted by  $s$  units, the partition function is transformed by

$$Z \rightarrow Z \cdot \exp \left( 2\pi i s \cdot \frac{d}{24} \right). \quad (2.24)$$

A topological invariant of framed, oriented three manifolds, together with a law for the behavior under change of framing, is more or less as good as a topological invariant of oriented three manifolds without a choice of framing.

Of course, all of the discussion in this section, and in particular (2.24), has been limited to the behavior at large  $k$ . In Sect. (4.5), we will see that the generalization of (2.24) to finite  $k$  is

$$Z \rightarrow Z \cdot \exp \left( 2\pi i s \cdot \frac{c}{24} \right), \quad (2.25)$$

with  $c$  being the central charge of two dimensional current algebra with symmetry group  $G$  at level  $k$ . It is well known that the large  $k$  limit of  $c$  is exactly  $d$ .

### *Moduli Spaces of Flat Connections*

There is still an important gap in the above discussion of the large  $k$  behavior. The formula (2.8) is really only valid if the determinants that appear are all non-zero. In fact, the flat connection  $A^{(\alpha)}$  determines a flat bundle  $E$ . The determinants in (2.8) are non-zero if and only if  $A^{(\alpha)}$  is such that the de Rham cohomology of  $M$ , with values in  $E$ , is zero. If  $H^1(M, E) \neq 0$ , then the flat connection  $A^{(\alpha)}$  is not isolated but lies on a moduli space  $\mathcal{S}$  of gauge inequivalent flat connections; and the proper evaluation of the path integral (2.1) leads not to the discrete sum (2.4) but to an integral on  $\mathcal{S}$ . If  $H^0(M, E)$  is not zero, then the fields  $\phi$ ,  $c$  and  $\bar{c}$  in the above treatment have zero modes, and the gauge fixing requires more care. It is plausible that by more careful study of the path integral, the large  $k$  contribution of arbitrary flat connections can be extracted without assumptions about  $H^*(M, E)$ . But we will not attempt this.

<sup>4</sup> The crucial factor of  $1/12$  in (2.22) reflects the discrepancy between the Chern character  $e^x = 1 + x^2/2 + \dots$  that appears in gauge theory index theorems and the  $\hat{A}$  genus  $(x/2)/\sinh(x/2) = 1 - x^2/24 + \dots$  that appears in gravitational index theorems

### Some Examples

We will later on determine the partition functions of some simple three manifolds, giving results that can be compared to large  $k$  computations. For  $S^2 \times S^1$ ,  $Z = 1$ , for any  $G$  and any  $k$ . For  $S^3$  and  $G = SU(2)$ , we will obtain the formula<sup>5</sup>

$$Z(S^3) = \sqrt{\frac{2}{k+2}} \sin\left(\frac{\pi}{k+2}\right). \quad (2.26)$$

Of course, on  $S^3$  the only flat connection is the trivial connection, for which (2.8) is not valid, since  $H^0(M, E) \neq 0$  in this case. For  $G = SU(2)$ , the behavior  $Z \sim k^{-3/2}$  in (2.26) is probably the general behavior of the contribution of the flat connection for homology spheres (on which the flat connection is isolated); it would be interesting to know how to obtain this behavior from path integrals. In Donaldson and Floer theory, the trivial connection, which has a negative formal dimension, is the cause of many subtleties. The vanishing of (2.26) in the classical limit of large  $k$  appears to be an interesting quantitative reflection of the “negative dimension” of the trivial connection.

### 2.1. Incorporation of Knots

We now wish to consider the large  $k$  behaviour in the presence of knots. For simplicity, we will limit ourselves to the case of  $S^3$ , and an abelian gauge group  $G = U(1)$ . Though the abelian gauge group is relatively trivial in the context of knot theory, it gives a quick and simple way to confirm the fact that the Chern-Simons action really does lead to topological invariants, and it also gives a simple context for explaining a technicality that is crucial in all that follows.

In the abelian theory, the gauge field is simply a one form  $A$  and the Lagrangian is

$$\mathcal{L} = \frac{k}{8\pi} \int_M e^{ijk} A_i \partial_j A_k. \quad (2.27)$$

We pick some circles  $C_a$  and some integers  $n_a$  [corresponding to representations of the gauge group  $U(1)$ ]. As always in this paper, we assume  $C_a$  does not intersect  $C_b$  for  $a \neq b$ . We wish to calculate the expectation value of the product

$$W = \prod_{a=1}^s \exp\left(in_a \int_{C_a} A\right) \quad (2.28)$$

with respect to the Gaussian measure determined by  $e^{i\mathcal{L}}$ . As was recently discussed by Polyakov (in a paper [23] in which he proposed to apply the Abelian Chern-Simons theory to high temperature superconductors), the result can be written in the form

$$\langle W \rangle = \exp\left(\frac{i}{2k} \sum_{a,b} n_a n_b \int_{C_a} dx^i \int_{C_b} dy^j \varepsilon_{ijk} \cdot \frac{(x-y)^k}{|x-y|^3}\right). \quad (2.29)$$

Here one has identified a region  $U$  of  $S^3$  containing the knots with a region of three dimensional Euclidean space, and  $x^i$ ,  $y^j$  are the Euclidean coordinates of  $U$

<sup>5</sup> The appearance of  $k+2$  in this formula is presumably an illustration of the  $k + c_2(G)/2$  in (2.17)

evaluated along the knots. For  $a \neq b$ , the integral in (2.29) is essentially the Gauss linking number, which can be written as

$$\Phi(C_a, C_b) = \frac{1}{4\pi} \int_{C_a} dx^i \int_{C_b} dy^j \varepsilon_{ijk} \frac{(x-y)^k}{|x-y|^3}. \quad (2.30)$$

As long as  $C_a$  and  $C_b$  do not intersect,  $\Phi(C_a, C_b)$  is a well defined integer; in fact, it is the most classic invariant in knot theory. Thus, if we could ignore the term  $a = b$ , we would have

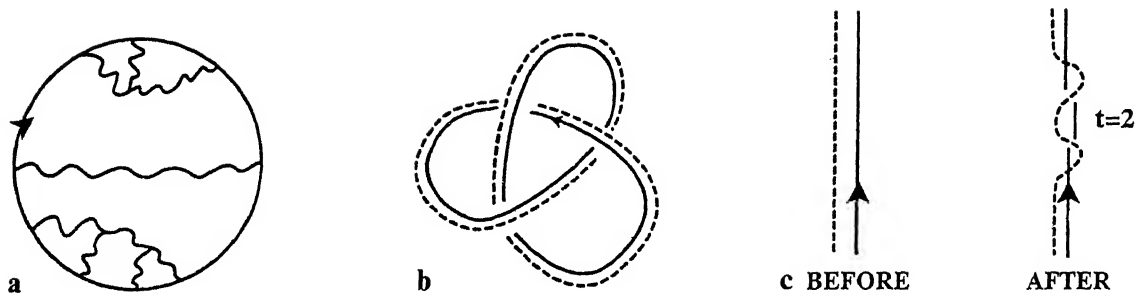
$$\langle W \rangle = \exp \left( \frac{2i\pi}{k} \sum_{a,b} n_a n_b \Phi(C_a, C_b) \right). \quad (2.31)$$

The appearance of the Gauss linking number illustrates the fact that the Chern-Simons theory does lead to topological invariants as we hope. But we have to worry about the term with  $a = b$ . This integral is ill-defined near  $x = y$ ; how do we wish to interpret it?

It is well known in knot theory that there is no natural and topologically invariant way to regularize the self-linking number of a knot. Polyakov in [23] used a regularization that is not generally covariant to get an answer that is interesting geometrically but not a topological invariant. We need a different approach for our present treatment in which general covariance is a primary goal. Though there is no completely invariant substitute for Polyakov's regularization, in the sense that there is no way to get a natural topological invariant from the integral in (2.29) or (2.30) with  $a = b$ , we cannot simply throw away the self-linking term and its non-abelian generalizations (which are sketched in Fig. 3a), since these terms are in fact not naturally zero. There is no reason to think that one could retain general covariance by dropping these terms. In the abelian theory, on a general three manifold  $M$ , on topological grounds the self-linking number can be a non-zero fraction, well-defined only modulo one. In such a case, it cannot be correct to set the self-linking number to zero, since it is definitely not zero. [Topologically, in such a situation, the self-linking number is well defined only modulo an integer, and this precision is definitely not good enough to evaluate (2.31).] In the non-abelian theory, we will get results later which amount to assigning definite, non-zero values to the non-abelian generalizations of the self-linking integral, so it would not be on the right track to try to throw these terms away.

Topologically, it is clear what data are needed to make sense of the self-linking of a knot  $C$ . One needs to give a "framing" of  $C$ ; this is a normal vector field along  $C$ . The idea is that by displacing  $C$  slightly in the direction of this vector field one gets a new knot  $C'$ , and it makes sense to calculate the linking number of  $C$  and  $C'$ . This can be defined as the self-linking number of the framed knot  $C$ . One can think of the framing as a thickening of the knot into a tiny ribbon bounded by  $C$  and  $C'$ ; this is how it is drawn in Fig. 3b. It is clear that the self-linking number defined this way depends not on the actual vector field used to displace  $C$  to  $C'$  but only on the topological class of this vector field; and indeed by a "framing" we mean only the topological class. Though a choice of framing gives a definition of the self-linking number of a knot  $C$ , it is clear that by picking a convenient framing of  $C$  one can





**Fig. 3a–c.** The self-linking integral is, in a non-abelian theory, the first in an infinite series of Feynman diagrams, with gauge fields emitted and absorbed by the same knot, as in **a**; these all pose similar problems. A topologically invariant but not uniquely determined regularization can be obtained by supposing that each knot is “framed”, as in **b**. In **c**, the framing is shifted by 2 units by making a 2-fold twist

get any desired answer for its self-linking number; as illustrated in Fig. 3c, a  $t$ -fold twist in the framing of  $C$  will change its self-linking by  $t$ .<sup>6</sup>

Physically, the role of the framing is that it makes possible what physicists would call a point-splitting regularization. This is defined as follows: when one has to do the self-linking integral in (2.29), one lets  $x$  run on  $C$  and  $y$  on  $C'$ . This gives a well-defined integral, though of course it depends on the framing. In this paper, we will assume, without proof, that the framing gives sufficient information to make possible a consistent point-splitting regularization of all the non-abelian generalizations of the self-linking integral, without further arbitrary choices. This question is, perhaps, comparable to the question of whether the non-abelian Chern-Simons action defines a sensible quantum theory in the first place (even without introducing Wilson lines as observables); neither of these questions will be tackled here.

Of course, if it were always possible to pick a canonical framing of knots, then we could pick this framing and hide the question. On  $S^3$ , there is a canonical framing of every knot; it is determined by asking that the self-linking number should be zero. (This makes the abelian linking integral zero, but not its non-abelian generalizations.) On general three manifolds, this cannot be done since the self-linking number may be ill-defined or may differ from an integer by a definite fraction (so that it does not vanish with any choice of framing). Even when the canonical framing does exist, it is not convenient to be restricted to using it, since natural operations (like the surgery we study in Sect. 4) may not preserve it.

In general, therefore, we give up on finding a natural choice, and simply pick some framing and proceed. It would be rather unpleasing if the “physical” results depended uncontrollably on the framing of knots. What saves the day is that although we cannot in general make a natural choice of the framing, we can state a general rule for how expectation values of Wilson lines change under a change of the framing. First of all, let us note that while, in general, there is no canonical zero in the set of possible framings of a knot in a three manifold, if one compares two framings they always differ by a definite integer, which is the relative twist in going

<sup>6</sup> The discussion should make it clear that the need to frame knots is analogous to the need to frame three manifolds, as found in the last section. This hopefully justifies the use of the same word “framing” in each case

around the knot (Fig. 3c). (That is, in general there is no natural way to count how many times the ribbon in Fig. 3b is twisted, but there is a natural local operation of adding  $t$  extra twists to this ribbon.) In the abelian theory, it is clear from (2.29) and (2.30) how the partition function transforms under a change of framing. If we shift the framing of the link  $C_a$  by  $t$  units, its self-linking number is increased by  $t$ , and the partition function is shifted by a phase

$$\langle W \rangle \rightarrow \exp(2\pi i t \cdot (n_a^2/k)) \cdot \langle W \rangle. \quad (2.32)$$

The nonabelian analog of that result will be derived in Sect. 5.1; the transformation law in the non-abelian case is

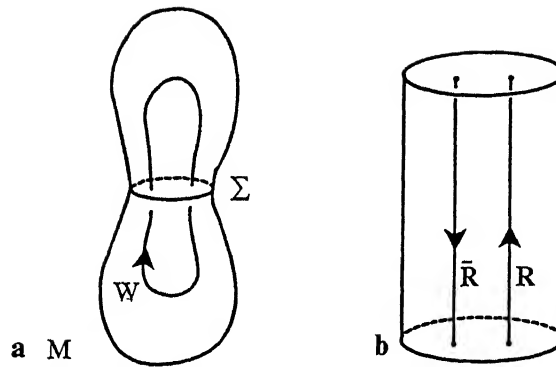
$$\langle W \rangle \rightarrow \exp(2\pi i t \cdot h) \langle W \rangle, \quad (2.33)$$

where  $h$  is the conformal weight of a certain primary field in  $1+1$  dimensional current algebra. This result, though it may seem rather technical, is a key ingredient enabling the Chern-Simons theory to work. It means that although we need to pick a framing for every link, because the self-linking integrals have no natural definition otherwise, there is no loss of information since we have a definite law for how the partition functions transform under change of framing.

Actually, it can be shown [13] that the structure of rational conformal field theory requires non-trivial monodromies. In the relationship that we will develop between the  $2+1$  dimensional Chern-Simons theory and rational conformal field theory in  $1+1$  dimensions, the need to frame all knots is the  $2+1$  dimensional analog of the monodromies that arise in  $1+1$  dimensions. [This will be clear in the derivation of (2.33).] Were it not for the seeming nuisance that knots must be framed to define the Wilson lines as quantum observables, one would end up proving that the Jones knot invariants were trivial.

An alternative description may make the physical interpretation of the framing of knots more transparent. A Wilson line can be regarded as the space-time trajectory of a charged particle. In  $2+1$  dimensions, it is possible for a particle to have fractional statistics, meaning that the quantum wave function changes by a phase  $e^{2\pi i \delta}$  under a  $2\pi$  rotation. (See [30] for a discussion of these issues.) If one wishes to compute a quantum amplitude with propagation of a particle of fractional statistics, it is not enough to specify the orbit of the particle; it is necessary to also count the number of  $2\pi$  rotations that the particle undergoes in the course of its motion. Equations (2.32) and (2.33) mean that the particles represented by Wilson lines in the Chern-Simons theory have fractional statistics with  $\delta = n_a^2/2k$  in the abelian theory or  $\delta = h$  in the non-abelian theory. This fractional statistics is the phenomenon claimed by Polyakov in [23], so in essence we agree with his substantive claim, though we prefer to exhibit this phenomenon in the context of a generally covariant regularization, where it appears in the behavior of Wilson lines under change of framing.

In this section, we have obtained some important evidence that the Chern-Simons theory can be regularized to give invariants of three manifolds and knots. We have also obtained the important insight that doing so requires picking a homotopy class of trivializations of the tangent bundle, and a “framing” of all knots. To actually solve the theory requires very different methods, to which we turn in the next section.



**Fig. 4a and b.** Cutting a three manifold  $M$  on an intermediate Riemann surface  $\Sigma$  is indicated in part a. Wilson lines  $W$  on  $M$  may pierce  $\Sigma$  and if so  $\Sigma$  comes with certain “marked points”, with representations attached. Locally, near  $\Sigma$ ,  $M$  looks like  $\Sigma \times R^1$ , indicated in part b

### 3. Canonical Quantization

The basic strategy for solving the Yang-Mills theory with Chern-Simons action on an arbitrary three manifold  $M$  is to develop a machinery for chopping  $M$  in pieces, solving the problem on the pieces, and gluing things back together. So to begin with we consider a three manifold  $M$ , perhaps with Wilson lines, as in Fig. 4a. We “cut”  $M$  along a Riemann surface  $\Sigma$ . Near the cut,  $M$  looks like  $\Sigma \times R^1$ , and our first step in learning to understand the theory on an arbitrary three manifold is to solve it on  $\Sigma \times R^1$ .

The special case of a three manifold of the form  $\Sigma \times R^1$  is tractable by means of canonical quantization. Canonical quantization on  $\Sigma \times R^1$  will produce a Hilbert space  $\mathcal{H}_\Sigma$ , “the physical Hilbert space of the Chern-Simons theory quantized on  $\Sigma$ ”.<sup>7</sup> These will turn out to be finite dimensional spaces, and moreover spaces that have already played a noted role in conformal field theory. In rational conformal field theories, one encounters the “conformal blocks” of Belavin, Polyakov, and Zamolodchikov. Segal has described these in terms of “modular functors” that canonically associate a Hilbert space to a Riemann surface, and has described in algebra-geometric terms a particular class of modular functors, which arise in current algebra of a compact group  $G$  at level  $k$  [16]. The key observation in the present work was really the observation that precisely those functors can be obtained by quantization of a three dimensional quantum field theory, and that this three dimensional aspect of conformal field theory gives the key to understanding the Jones polynomial.

<sup>7</sup> It is conventional in physics to call vector spaces obtained in this fashion “Hilbert spaces”, and we will follow this terminology. In fact, the claim that comes most naturally from path integrals and that we will actually use is only that  $\mathcal{H}_\Sigma$  is a vector space canonically associated with  $\Sigma$ , and exchanged with its dual when the orientation of  $\Sigma$  is reversed. However, a Hilbert space structure is natural in the Hamiltonian viewpoint, and in the particular problem we are considering here, an inner product on  $\mathcal{H}_\Sigma$  is important in more delicate aspects of conformal field theory; such an inner product gives a “metric on the flat vector bundle” in the language of Friedan and Shenker [31]. According to Segal [16],  $\mathcal{H}_\Sigma$  in fact has a canonical projective Hilbert space structure

Actually, the general situation that must be studied is that in which possible Wilson lines on  $M$  are “cut” by  $\Sigma$ , as in the figure. In this case  $\Sigma$  is presented with finitely many marked points  $P_1, \dots, P_k$ , with a  $G$  representation  $R_i$  assigned to each  $P_i$  (since each Wilson line has an associated representation). To this data – an oriented topological surface with marked points, and for each marked point a representation of  $G$  – we wish to associate a vector space. This is also the general situation that arises in conformal field theory – the marked points are points at which operators with non-vacuum quantum numbers have been inserted. If one reverses the orientation of  $\Sigma$  (and replaces the representations  $R_i$  associated with the marked points with their complex conjugates) the vector space  $\mathcal{H}_\Sigma$  must be replaced with its dual.

*The Canonical Formalism.* At first sight, (1.3) might look like a typically intractable nonlinear quantum field theory, but this is far from being so. Working on  $\Sigma \times R^1$ , it is very natural to choose the gauge  $A_0 = 0$  (with  $A_0$  being the component of the connection in the  $R^1$  direction). In this gauge we immediately see that the Lagrangian becomes quadratic. It reduces to

$$\mathcal{L} = \frac{k}{8\pi} \int dt \int_\Sigma \varepsilon^{ij} \text{Tr} A_i \frac{d}{dt} A_j. \quad (3.1)$$

For the time being we will ignore extra complications due to Wilson lines that may be present on  $\Sigma \times R^1$ . From (3.1) we may deduce the Poisson brackets,<sup>8</sup>

$$\{A_i^a(x), A_j^b(y)\} = \frac{4\pi}{k} \cdot \varepsilon_{ij} \delta^{ab} \delta^2(x - y). \quad (3.2)$$

Before rushing ahead to quantize these commutation relations, we should remember that the system is subject to a “Gauss law” constraint, which is  $\delta \mathcal{L} / \delta A_0 = 0$ , or (ignoring the Wilson lines)

$$\varepsilon^{ij} F_{ij}^a = 0. \quad (3.3)$$

This constraint equation is nonlinear (since  $F$  contains a quadratic term), and – as (3.1) is certainly a free theory – this nonlinearity is what remains of the underlying nonlinearity of (1.3).

In quantum field theory, one very often quantizes first and then imposes the constraints. The situation that we are considering here is a situation in which it is far more illuminating to first impose the constraints and then quantize. For the phase space  $\mathcal{M}_0$  of connections  $A_i^a(x)$  without the constraints is an infinite dimensional phase space; imposing the constraints will reduce us to a rather subtle but eminently finite dimensional phase space  $\mathcal{M}$ . The problem that faces us here, of reducing from  $\mathcal{M}_0$  to  $\mathcal{M}$  by imposing the constraints (3.2), has been studied before – and has proved to have extremely rich properties – in the work of Atiyah and Bott on equivariant Morse theory, two dimensional Yang-Mills theory, and

<sup>8</sup> This is a typical problem in which it is not appropriate to “introduce canonical momenta”. The purpose of introducing such variables is to reexpress a given Lagrangian in a form which is first order in time derivatives, but (3.1) is already first order in time derivatives. The variables in (3.1) are already canonically conjugate, as indicated in the following equation

the moduli space of holomorphic vector bundles [33]. In our present investigation, this familiar problem appears from a novel three dimensional vantage point.

It is necessary to recall the nature of constraint equations in classical physics. The constraints (3.2) are functions that should vanish, but they also generate gauge transformations via Poisson brackets. Imposing the constraints means two things classically: First, we restrict ourselves to values of the canonical variables for which the constraint functions vanish; and second, we identify two solutions of the constraint equations if they differ by a gauge transformation. In the case at hand, the first step means that we should consider only “flat connections”, that is, connections for which  $F_{ij}^a = 0$ . The second step means that we identify two flat connections if they differ by a gauge transformation. Taking the two steps together, we see that the physical phase space, obtained by imposing the constraints (3.2), is none other than the moduli space of flat connections on  $\Sigma$ , modulo gauge transformations. Such flat connections are completely characterized by the “Wilson lines”, that is, the holonomies around non-contractible loops on  $\Sigma$ . A simple count of parameters shows that on a Riemann surface of genus  $g > 1$ , the moduli space  $\mathcal{M}$  of flat connections modulo gauge transformations has dimension  $(2g - 2) \cdot d$ , where  $d$  is the dimension of the group  $G$ .

The topology of  $\mathcal{M}$  is rather intricate (and this was in fact the main subject of interest in [33]). On general grounds  $\mathcal{M}$  inherits a symplectic structure (that is, a structure of Poisson brackets) from the symplectic structure present on  $\mathcal{M}_0$  before imposing the constraints.  $\mathcal{M}$  is a compact space (with some singularities), and in particular its volume with the natural symplectic volume element is finite. Since in quantum mechanics there is one quantum state per unit volume in classical phase space, the finiteness of the volume of  $\mathcal{M}$  means that the quantum Hilbert spaces will be finite dimensional. We would like to determine them.

### 3.1. The Holomorphic Viewpoint

Quantization of classical mechanics is usually carried out by separating the canonical variables into “coordinates”,  $q^i$ , which are a maximal set of real commuting variables, and “momenta”,  $p^j$ , which are conjugate to the  $q^i$ . The quantum Hilbert space is then the space  $\mathcal{H}$  of square integrable functions of the  $q^i$ .

Such a scheme definitely requires a noncompact phase space of infinite volume, since – though the  $q^i$  may take values in a compact space – the  $p^j$  are definitely unbounded. Accordingly, the space  $\mathcal{H}$  is infinite dimensional.

Quantizing a compact, finite volume phase space, such as the moduli space  $\mathcal{M}$  of flat connections modulo gauge transformations, is quite a different kind of problem. It has no known general solution, but there is one important class of cases in which there is a natural notion of quantization. This arises in the case in which  $\mathcal{M}$  is a Kähler manifold, and the symplectic structure on  $\mathcal{M}$  is the curvature form that represents the first Chern class of a holomorphic line bundle  $L$  endowed with some metric. In this case, one carries out quantization not by separating the variables in phase space into “coordinates” and “momenta”,  $q$ ’s and  $p$ ’s, but by separating them into holomorphic and anti-holomorphic degrees of freedom, essentially  $z \sim q + ip$  and  $\bar{z} \sim q - ip$ . The quantum Hilbert space  $\mathcal{H}$  is then a suitable space of holomorphic “functions”. More exactly,  $\mathcal{H}$  is the space of

holomorphic sections of the line bundle  $L$ . If  $\mathcal{M}$  is compact, this latter space will be finite dimensional. In our problem, with  $\mathcal{M}$  being the moduli space of flat connections modulo gauge transformations on an oriented smooth surface  $\Sigma$ , is there a natural Kähler structure on  $\mathcal{M}$ ? The answer is crucial for all that follows. There is not quite a *natural* Kähler structure on  $\mathcal{M}$ , but there is a natural way to obtain such structures. Once one picks a complex structure  $J$  on  $\Sigma$ , the moduli space  $\mathcal{M}$  of flat connections can be given a new interpretation – it is the moduli space of stable holomorphic  $G_{\mathbb{C}}$  bundles on  $\Sigma$  which are topologically trivial ( $G_{\mathbb{C}}$  is the complexification of the gauge group  $G$ ). Let us refer to the latter space as  $\mathcal{M}_J$ .  $\mathcal{M}_J$  is naturally a complex Kähler (and in fact projective algebraic) variety. Upon picking a linear representation of  $G$  (for our purposes it is convenient to pick a representation with the smallest value of the quadratic Casimir operator, e.g. the  $N$  dimensional representation of  $SU(N)$  or the adjoint representation of  $E_8$ ), and passing from a principal  $G_{\mathbb{C}}$  bundle to the associated vector bundle, we can think of  $\mathcal{M}_J$  as the moduli space of a certain family of holomorphic vector bundles. For  $G = SU(N)$ ,  $\mathcal{M}_J$  is simply the moduli space of all stable rank  $N$  holomorphic vector bundles of vanishing first Chern class.

The symplectic form on  $\mathcal{M}$  that appears in (3.1) or (3.2) *without* picking a complex structure on  $\Sigma$  has a very special interpretation in holomorphic terms once we *do* pick such a complex structure. Let us recall the notion [34] of the determinant line bundle of the  $\bar{\partial}$  operator. The  $\bar{\partial}$  operator on  $\Sigma$  can be “twisted” by any holomorphic vector bundle.  $\mathcal{M}_J$  parametrizes a family of holomorphic vector bundles on  $\Sigma$ , and thus it can be regarded as parametrizing a family of  $\bar{\partial}$  operators. Taking the determinant line gives a line bundle  $L$  over the base space  $\mathcal{M}_J$  of this family. Furthermore [34], the Dirac determinant gives a natural metric on  $L$ , and the first Chern class of  $L$ , computed with this metric, is precisely the symplectic form that appears in (3.1) or (3.2), provided  $k = 1$ . For general  $k$ , the symplectic form that appears in (3.1) or (3.2) represents the first Chern class of the  $k^{\text{th}}$  power of the determinant line bundle.<sup>9</sup>

Thus, all of the conditions are met for a straightforward quantization of (3.1), taking into account the constraints (3.3). The constraints mean that the classical space to be quantized is the moduli space  $\mathcal{M}$  of flat connections. Picking an arbitrary complex structure  $J$  on  $\Sigma$ ,  $\mathcal{M}$  becomes a complex manifold, and the symplectic form of interest represents the first Chern class of  $L^{\otimes k}$ , the  $k^{\text{th}}$  tensor power of the determinant line bundle. The quantum Hilbert space  $\mathcal{H}_{\Sigma}$  is thus the space of global holomorphic sections of  $L^{\otimes k}$ .

### 3.2. A Flat Vector Bundle on Moduli Space

This gives an answer to the problem of canonically quantizing the Chern-Simons theory on  $\Sigma \times R^1$ , but a crucial point now requires discussion.

<sup>9</sup> This description is valid for the gauge group  $G = SU(N)$ , but in general the following modification is needed. For groups other than  $SU(N)$  the determinant line bundle  $L$  is not the fundamental line bundle on  $\mathcal{M}$  but a tensor power thereof. For instance, for  $G = E_8$ , there is a line bundle  $L'$  with  $(L')^{\otimes 30} \simeq L$ . It is then  $L'$  whose first Chern class corresponds to (3.1) or (3.2) with  $k = 1$ .

Quantizing (3.1), with the constraints (3.3), is a problem that can be naturally asked whenever one is given an oriented smooth surface  $\Sigma$ . Beginning with a generally covariant Lagrangian in three dimensions, we were led to this problem in a context in which it was not natural to assume any metric or complex structure on  $\Sigma$ . However, to solve the problem and construct  $\mathcal{H}_\Sigma$ , it was very natural to pick a complex structure  $J$  on  $\Sigma$ . Thus, our description of  $\mathcal{H}_\Sigma$  depends on the choice of  $J$ , and what we have called  $\mathcal{H}_\Sigma$  might perhaps be better called  $\mathcal{H}_\Sigma^{(J)}$ . As  $J$  varies, the  $\mathcal{H}_\Sigma^{(J)}$  vary holomorphically with  $J$ , and thus we could interpret this object as a holomorphic vector bundle on the moduli space of complex Riemann surfaces. But since  $\mathcal{H}_\Sigma^{(J)}$  is the answer to a question that depends on  $\Sigma$  and not on  $J$ , we would like to believe that likewise the  $\mathcal{H}_\Sigma^{(J)}$  canonically depend only on  $\Sigma$  and not on  $J$ . The assertion that the  $\mathcal{H}_\Sigma^{(J)}$  are canonically independent of  $J$ , and depend only on  $\Sigma$ , is the assertion that the vector bundle on moduli space given by the  $\mathcal{H}_\Sigma^{(J)}$  has a canonical flat connection that permits one to identify the fibers. Such “flat vector bundles on moduli space” first entered in conformal field theory somewhat implicitly in the differential equations of Belavin, Polyakov, and Zamolodchikov [32]. They were discussed much more explicitly by Friedan and Shenker [31], who proposed that they would play a pivotal role in conformal field theory, and they have been prominent in subsequent work such as [12, 13]. At least in one important class of examples, we have just met a natural origin of “flat vector bundles on moduli space”. The problem “quantize the Chern-Simons action” can be posed without picking a complex structure, so the answer is naturally independent of complex structure and thus gives a “flat bundle on moduli space”. The particular flat bundles on moduli space that we get this way are those that Segal has described [16] in connection with conformal field theory; Segal also rigorously proved the flatness, which is explained somewhat heuristically by the physical argument sketched above. (Because of the conformal anomaly, this bundle has only a projectively flat connection, with the projective factor being canonically odd under reversal of orientation.)

The role of these flat bundles in conformal field theory is as follows. If one considers current algebra on a Riemann surface, with a symmetry group  $G$ , at “level”  $k$ , then one finds that in genus zero the Ward identities uniquely determine the correlation functions for descendants of the identity operator, but this is not so in genus  $\geq 1$ . On a complex Riemann surface  $\Sigma$  of genus  $\geq 1$ , the space of solutions of the Ward identities for descendants of the identity is a vector space  $\hat{\mathcal{H}}_\Sigma$ , which might be called the “space of conformal blocks”. Segal calls the association  $\Sigma \rightarrow \hat{\mathcal{H}}_\Sigma$  a “modular functor”, and has given an algebra-geometric description of the modular functors that arise in current algebra. In quantizing the Chern-Simons theory we have exactly reproduced this description! This is then the secret of the relation between current algebra in  $1+1$  dimensions and Yang-Mills theory in  $2+1$  dimensions: the space of conformal blocks in  $1+1$  dimensions are the quantum Hilbert spaces obtained by quantizing a  $2+1$  dimensional theory. It would take us too far afield to explain here the algebra-geometric description of the space of conformal blocks. Suffice it to say that when one tries to use the Ward identities of current algebra to uniquely determine the correlation functions of descendants of the identity on a curve  $\Sigma$  of genus  $\geq 1$ , one meets an obstruction which involves the existence of non-trivial holomorphic vector bundles on  $\Sigma$ ; the



Ward identities reduce the determination of the correlation functions to the choice of a holomorphic section of  $L^{\otimes k}$  over the moduli space of bundles.

It seems appropriate to conclude this discussion with some remarks on the formal properties of the association  $\Sigma \rightarrow \mathcal{H}_\Sigma$ . It is good to first think of the functor  $\Sigma \rightarrow H^1(\Sigma, R)$  which to a Riemann surface  $\Sigma$  associates its first de Rham cohomology group. This functor is defined for every smooth surface  $\Sigma$ , independent of complex structure. A diffeomorphism of  $\Sigma$  induces a linear transformation on  $H^1(\Sigma, R)$ , so  $H^1(\Sigma, R)$  furnishes in a natural way a representation of the mapping class group. The formal properties of the functors  $\Sigma \rightarrow \mathcal{H}_\Sigma$  that come by quantizing the Chern-Simons theory are quite analogous. Though a complex structure  $J$  on  $\Sigma$  is introduced to construct  $\mathcal{H}_\Sigma$ , the existence of a natural projectively flat connection on the moduli space of complex structures permits one locally to (projectively) identify the various  $\mathcal{H}_\Sigma^{(J)}$  and forget about the complex structure. One might think that the global monodromies of the flat connection on moduli space would mean that globally one could not forget the complex structure, but this is not so; these monodromies just correspond to an action of the purely topological mapping class group, so that the formal properties of  $\mathcal{H}_\Sigma$  are just like those of  $H^1(\Sigma, R)$ .

### 3.3. Inclusion of Wilson Lines

So far, we have discussed the quantization of the Chern-Simons theory on a Riemann surface  $\Sigma$  *without* Wilson lines. Now we wish to include the Wilson lines, which, as in Fig. 4b, pierce  $\Sigma$  in some points  $P_i$ ; associated with each such point is a representation  $R_i$ . Quantizing the Chern-Simons theory in the presence of the Wilson lines should give a Hilbert space  $\mathcal{H}_{\Sigma; P_i, R_i}$  that is canonically associated with the oriented surface  $\Sigma$  together with the choice of  $P_i$  and  $R_i$ .

It is pretty clear what problem in conformal field theory this should correspond to. Instead of simply considering correlation functions of the descendants of the identity, we should consider in the conformal field theory primary fields transforming in the  $R_i$  representations of  $G$ . With these fields (or their descendants) inserted at points  $P_i$  on  $\Sigma$ , one gets in conformal field theory a more elaborate space  $\hat{\mathcal{H}}_{\Sigma; P_i, R_i}$  of conformal blocks. Again, there is an algebra-geometric description of this space [16], and this is what we should expect to recover by quantizing the Chern-Simons theory in the presence of the Wilson lines.

I will now briefly sketch how this works out, deferring a fuller treatment for another occasion. First of all, the Wilson lines correspond to static non-abelian charges which show up as extra terms in the constraint equations. So (3.3) is replaced by

$$\frac{k}{8\pi} \varepsilon^{ij} F_{ij}^a(x) = \sum_{s=1}^r \delta^2(x - P_s) T_{(s)}^a, \quad (3.4)$$

where  $P_s$ ,  $s=1 \dots r$  are the points at which static external charges have been placed, and  $T_{(s)}^a$ ,  $a=1 \dots \dim G$  are the group generators associated with the external charges. Now, a naive attempt to quantize (3.1) with the generalized constraints (3.4) would run into extremely unpleasant difficulties. One could try to quantize first and then impose the constraints, but this is difficult to see through



even in the absence of the external charges. Alternatively, one can try to impose the constraints at the classical level and then quantize, as we did above. But it is hard to make sense of (3.4) as constraints in the classical theory; the solution  $A_i^a$  of (3.4) cannot be an ordinary  $c$ -number connection, since non-commuting operators appear on the right-hand side. It is clear that to solve (3.4),  $A_i^a$  would have to be some sort of “ $q$ -number connection”, whose holonomy would presumably be an element of a “quantum group”, not an ordinary classical group. Indeed, it seems likely that the theory of quantum groups [35] can be considered to arise in this way.

However, there is a much better way to quantize the Chern-Simons theory with static charges. We certainly wish to impose (3.4) at the classical level. This cannot be done directly, since on the right-hand side there appear quantum operators. A useful point of view is the following. A representation  $R_i$  of a group  $G$  should be seen as a quantum object. This representation should be obtained by quantizing a classical theory. The Borel-Weil-Bott theorem gives a canonical way to exhibit for every irreducible representation  $R$  of a compact group  $G$  a problem in classical physics, with  $G$  symmetry, such that the quantization of this classical problem gives back  $R$  as the quantum Hilbert space. One introduces the “flag manifold”  $G/T$ , with  $T$  being a maximal torus in  $G$ , and for each representation  $R$  one introduces a symplectic structure  $\omega_R$  on  $G/T$ , such that the quantization of the classical phase space  $G/T$ , with the symplectic structure  $\omega_R$ , gives back the representation  $R$ . Many aspects of representation theory find natural explanations by thus regarding representations of groups as quantum objects that are obtained by quantization of classical phase spaces.

In the problem at hand, this point of view can be used to good effect. We extend the phase space  $\mathcal{M}_0$  of  $G$  connections on  $\Sigma$  by including at each marked point  $P_i$  a copy of  $G/T$ , with the symplectic structure appropriate to the  $R_i$  representation. The quantum operators  $T_{(i)}^a$  that appear on the right of (3.4) can then be replaced by the classical functions on  $G/T$  whose quantization would give back the  $T_{(i)}^a$ . The constraints (3.4) then make sense as classical equations, and the analysis can be carried out just as we did without marked points, though the details are a bit longer. Suffice it to say that after imposing the classical constraints, one gets a finite dimensional phase space  $\mathcal{M}_{P_i, R_i}$  that incorporates the static charges; a point on this space is a flat  $G$  connection on  $\Sigma$  with a reduction of structure group to  $T$  at the points  $P_i$ . Upon picking an arbitrary conformal structure on  $\Sigma$ , this phase space can be quantized. In this way one gets exactly Segal’s description of the space of conformal blocks in current algebra in a general situation with primary fields in the  $R_i$  representation inserted at the points  $P_i$ . (In current algebra at level  $k$ , one only permits certain representations, the “integrable ones”. If one formally tries to include other representations, the Ward identities show that they decouple [36]. According to Segal, the analogous statement in algebraic geometry is that the appropriate line bundle over  $\mathcal{M}_{P_i, R_i}$  has no non-zero holomorphic sections unless the  $R_i$  all correspond to integrable representations. For the Chern-Simons theory, this means that unless the representations  $R_i$  are all integrable, the zero vector is the only vector in the physical Hilbert space.)

Finally, let us note that the Borel-Weil-Bott theorem should not be used simply as a tool in quantization. It should be built into the three dimensional description.

One should use the theorem to replace the Wilson lines (1.5) that appear in (1.6) with a functional integral over maps of the circle  $S$  into  $G/T$  (or actually an integral over sections of a  $G/T$  bundle, twisted by the restriction to  $S$  of the  $G$ -bundle  $E$ ). This gives a much more unified formalism.

### 3.4. The Riemann Sphere with Marked Points

The above description may seem a little bit dense, and we will supplement it by giving a simple intuitive description of the physical Hilbert space  $\mathcal{H}_{\Sigma; R_i, P_i}$  in the important case of genus zero. Let  $\Sigma$  be an oriented surface of genus zero, with static charges in the  $R_i$  representation at points  $P_i$ . Let us consider the case of very large  $k$ . Now, the gauge coupling in (1.3) is of order  $1/k$ , so for large  $k$  we are dealing with very weak coupling. Rather naively, one might believe that for extremely weak coupling the physical Hilbert space is the same as it would be if the charges were not coupled to gauge fields. If so, the physical Hilbert space would be simply the tensor product  $\mathcal{H}_0 = \otimes_i R_i$  of the Hilbert spaces  $R_i$  of the individual charges. However, there is a key error here. No matter how weak the gauge coupling may be, we must remember that in a closed universe the total charge must be zero (since the electric flux has nowhere to go). The total charge being zero means in a nonabelian theory that all of the charges together must be coupled to the trivial representation of  $G$ . So the physical Hilbert space, for large  $k$ , is precisely the  $G$ -invariant subspace of  $\mathcal{H}_0$ , or

$$\mathcal{H} = \text{Inv}(\otimes_i R_i). \quad (3.5)$$

This is a familiar answer in conformal field theory for the space of conformal blocks obtained, in the large  $k$  limit, in coupling representations  $R_i$ . Considerations of conformal field theory also show that for finite  $k$  the correct answer is always a subspace of (3.5). The most important modification of (3.5) that arises for finite  $k$  (and is explained algebra-geometrically in [16]) is that  $\mathcal{H}$  is zero unless the  $R_i$  correspond to integrable representations of the loop group; in what follows a restriction to such representations is always understood.

Now we consider some important special cases.

(i) For the Riemann sphere with no marked points, the Hilbert space is one dimensional. This is well known in conformal field theory – for descendants of the identity on the Riemann sphere, there is only one conformal block.

(ii) For the Riemann sphere with one marked point in a representation  $R_i$ , the Hilbert space is one dimensional if  $R_i$  is trivial, and zero dimensional otherwise.

(iii) For the Riemann sphere with two marked points with representations  $R_i$  and  $R_j$ , the Hilbert space is one dimensional if  $R_j$  is the dual of  $R_i$  (so that there is an invariant in  $R_i \otimes R_j$ ) and zero dimensional otherwise. Again, this is well known in conformal field theory.

(iv) For the Riemann sphere with three marked points in representations  $R_i$ ,  $R_j$ , and  $R_k$ , the dimension of  $\mathcal{H}$  is the number  $N_{ijk}$  for which Verlinde has proposed [12] and Moore and Seiberg have proved [13] rather striking properties. Here,  $N_{ijk}$  may in general be less than its large  $k$  limit which is the dimension of (3.5).

(v) From the results of Verlinde, the dimensions of the physical Hilbert spaces for an arbitrary collection of marked points on  $S^2$  can be determined from a knowledge of the  $N_{ijk}$ . But let us consider a particularly important special case.

Suppose that there are four external charges, and that the representations are  $R$ ,  $R$ ,  $\bar{R}$ , and  $\bar{R}$ . If the decomposition of  $R \otimes R$  is

$$R \otimes R = \bigoplus_{i=1}^s E_i, \quad (3.6)$$

with the  $E_i$  being distinct irreducible representations of  $G$ , then the physical Hilbert space  $\mathcal{H}$  at large  $k$  will be  $s$  dimensional, since the possible invariants in  $R \otimes R \otimes \bar{R} \otimes \bar{R}$  are uniquely fixed by giving the representation to which  $R \otimes R$  is coupled. (For small  $k$  the dimension of  $\mathcal{H}$  might be less than  $s$ .) In understanding the knot polynomials, an important special case is that in which  $G$  is  $SU(N)$  and  $R$  is the defining  $N$  dimensional representation. In that case,  $s = 2$  and the physical Hilbert space is two dimensional (except for  $k = 1$  where it is one dimensional).

#### 4. Calculability

Our considerations so far may have seemed somewhat abstract, and we would now like to show that in fact these considerations can actually be used to calculate things. As an introduction to the requisite ideas, we will first deduce a certain theoretical principle that is of great importance in its own right.

Consider, as in Fig. 5a, a three manifold  $M$  which is the connected sum of two three manifolds  $M_1$  and  $M_2$ , joined along a two sphere  $S^2$ . There may be knots in  $M_1$  or  $M_2$ , but if so they do not pass through the joining two sphere. If for every three manifold  $X$  we denote the partition function or Feynman path integral (1.6) as  $Z(X)$ , then we wish to deduce the formula

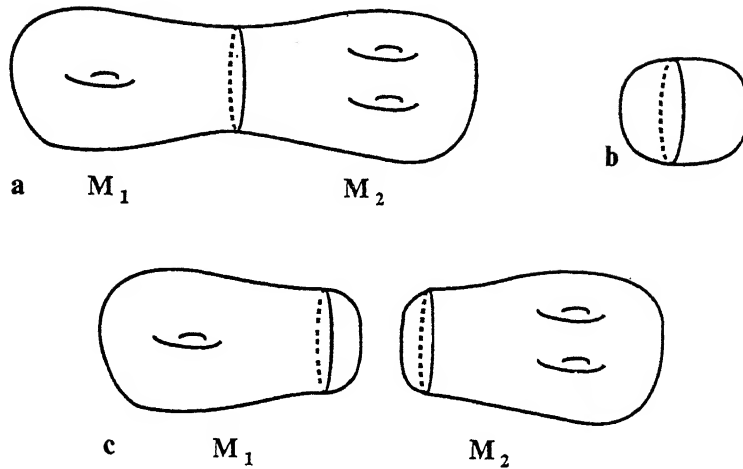
$$Z(M) \cdot Z(S^3) = Z(M_1) \cdot Z(M_2) \quad (4.1)$$

[it being understood that  $Z(S^3)$  denotes the partition function of a three sphere that contains no knots]. This can be rewritten

$$\frac{Z(M)}{Z(S^3)} = \frac{Z(M_1)}{Z(S^3)} \cdot \frac{Z(M_2)}{Z(S^3)}. \quad (4.2)$$

In some special cases, (4.2) is equivalent or closely related to known formulas. If  $M_1$  and  $M_2$  are copies of  $S^3$  with knots in them, then the ratios appearing in (4.2) turn out to be the knot invariants that appear in the Jones theory, and (4.2) expresses the fact that these invariants are multiplicative when one takes the disjoint sum of knots. If  $M_1$  and  $M_2$  are arbitrary three manifolds without knots, then (in view of our discussion in Sect. 2) (4.2) is closely related to the multiplicativity of Reidemeister and Ray-Singer torsion under connected sums.

So let us study Fig. 5a using the general ideas of quantum field theory. On the left of this figure, we see a three manifold  $M_1$  with boundary  $S^2$ . According to the general ideas of quantum field theory, one associates a “physical Hilbert space”  $\mathcal{H}$  with this  $S^2$ ; as we have seen in the last section, it is one dimensional. The Feynman path integral on  $M_1$  determines a vector  $\chi$  in  $\mathcal{H}$ . Likewise, on the right of



**Fig. 5a–c.** In **a** is sketched a three manifold  $M$  which is the connected sum of two pieces  $M_1$  and  $M_2$ , joined along a sphere  $S^2$ . Similarly, a three sphere  $S^3$  can be cut along its equator, as in **b**. Cutting both  $M$  and  $S^3$  as indicated in **a** and **b**, the pieces can be rearranged into the *disconnected* sum of  $M_1$  and  $M_2$ , as in **c**

Fig. 5a we see a three manifold  $M_2$  whose boundary is the same  $S^2$  with opposite orientation; its Hilbert space  $\mathcal{H}'$  is canonically the dual of  $\mathcal{H}$ . The path integral on  $M_2$  determines a vector  $\psi$  in  $\mathcal{H}'$ , and according to the general ideas of quantum field theory, the partition function of the connected sum  $M$  is

$$Z(M) = (\chi, \psi). \quad (4.3)$$

The symbol  $(\chi, \psi)$  denotes the natural pairing of vectors  $\chi \in \mathcal{H}$ ,  $\psi \in \mathcal{H}'$ . We cannot evaluate (4.3), since we do not know  $\chi$  or  $\psi$ . Instead, let us consider some variations on this theme. The two sphere  $S^2$  that separates the two parts of Fig. 5a could be embedded in  $S^3$  in such a way as to separate  $S^3$  into two three balls  $B_L$  and  $B_R$ . The path integrals on  $B_L$  and  $B_R$  would give vectors  $v$  and  $v'$  in  $\mathcal{H}$  and  $\mathcal{H}'$ , and the same reasoning as led to (4.3) gives

$$Z(S^3) = (v, v'). \quad (4.4)$$

Again, we do not know  $v$  or  $v'$  and cannot evaluate (4.4). But we can say the following. As  $\mathcal{H}$  is one dimensional,  $v$  is a multiple of  $\chi$ ; likewise, since  $\mathcal{H}'$  is one dimensional,  $v'$  is a multiple of  $\psi$ . It is then a fact of one dimensional linear algebra that

$$(\chi, \psi) \cdot (v, v') = (\chi, v') \cdot (v, \psi). \quad (4.5)$$

The two terms on the right-hand side of (4.5) are respectively  $Z(M_1)$  and  $Z(M_2)$ , as we see in Fig. 5c. So (4.5) is equivalent to the desired result (4.1).

One may wonder what is the mysterious object  $Z(S^3)$  that is so prominent in (4.1). Can it be set to one? Actually, the axioms of quantum field theory are strong enough so that the value of  $Z(S^3)$  is uniquely determined and cannot be postulated arbitrarily; as we will see later it can be calculated from the theory of affine Lie algebras. For  $G = SU(2)$  the formula has been given in (2.26).

As a special case of (4.1), pick  $s$  irreducible representations of  $G$ , say  $R_1, \dots, R_s$ , and consider a link in  $S^3$  that consists of  $s$  unlinked and unknotted circles  $C_i$ , with

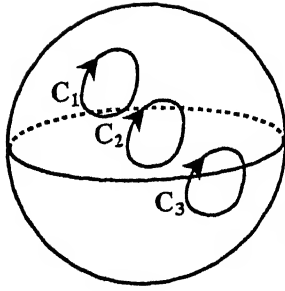


Fig. 6. A three sphere with 3 unlinked and unknotted circles  $C_i$ , associated with representations  $R_1 \dots R_3$ . The figure can be cut in various ways to separate the circles

one of the  $R_i$  associated with each circle. This is indicated in Fig. 6. Denote the partition function of  $S^3$  with this collection of Wilson lines as  $Z(S^3; C_1, \dots, C_s)$  (the representations  $R_i$  being understood). Then by cutting the figure to separate the circles, and repeatedly using (4.1), we learn that

$$\frac{Z(S^3; C_1, \dots, C_s)}{Z(S^3)} = \prod_{k=1}^s \frac{Z(S^3; C_k)}{Z(S^3)}. \quad (4.6)$$

If we introduce the normalized expectation value of a link  $L$ , defined by  $\langle L \rangle = Z(S^3; L)/Z(S^3)$ , then (4.6) becomes

$$\langle C_1 \dots C_s \rangle = \prod_k \langle C_k \rangle \quad (4.7)$$

for an arbitrary collection of unlinked, unknotted Wilson lines on  $S^3$ .

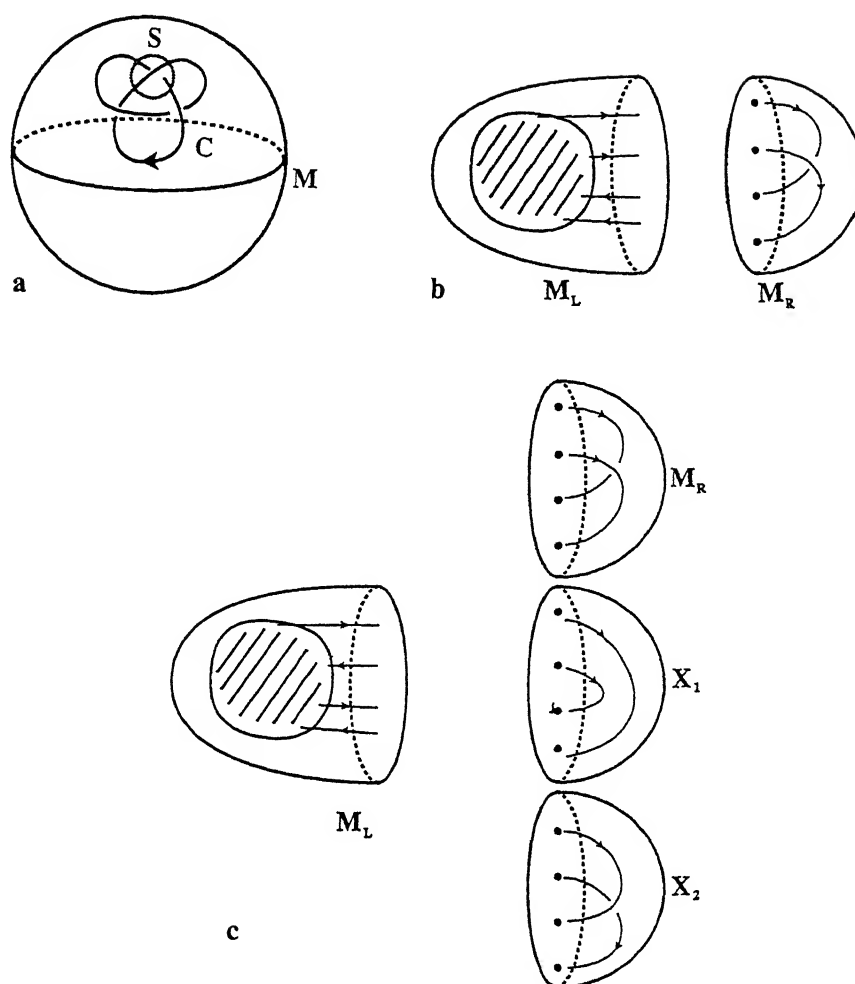
In knot theory there is another notion of connected sum, the “connected sum of links”. The Jones invariants also have a simple multiplicative behavior under this operation, as we will sketch briefly at the end of Sect. 4.5.

#### 4.1. Knots in $S^3$

We will now describe the origin of the “skein relation” which can be taken as the definition of the knot polynomials for knots on  $S^3$ . (A special case of the skein relation was first used by Conway in connection with the Alexander polynomial.)

Consider a link  $L$  on a general three manifold  $M$ , as indicated in Fig. 7a. The components of the link are associated with certain representations of  $G$ , and we wish to calculate the Feynman path integral (1.6), which we will denote as  $Z(L)$  (with the representations understood). We will evaluate it by deducing an algorithm for unknotting knots. If the lines in Fig. 7 could pass through each other unimpeded, all knots could be unknotted. As it is, this is prevented by some unfortuitous crossings, such as the one circled in the figure. Let us draw a small sphere about this crossing, cut it out, and study it more closely. This cuts  $M$  into two pieces, which after rearrangement are shown in Fig. 7b as a complicated piece  $M_L$  shown on the left of the figure and a simple piece  $M_R$  shown on the right.  $M_R$  consists of a three ball with boundary  $S^2$ ; on this boundary there are four marked points that are connected by two lines in the interior of the ball.

To make the discussion concrete, let us suppose that the gauge group is  $G = SU(N)$  and that the Wilson lines are all in the defining  $N$  dimensional representation of  $SU(N)$ , which we will call  $R$ . Then, as we saw at the end of the



**Fig. 7a–c.** A link  $C$  on a general three manifold  $M$  is sketched in a. A small sphere  $S$  has been drawn about an inconvenient crossing; it cuts  $M$  into a simple piece (the interior of  $S$ ) and a complicated piece. In b, the picture is rearranged to exhibit the cutting of  $M$  more explicitly; the two pieces now appear on the left and right as  $M_L$  (the complicated piece whose details are not drawn) and  $M_R$  (the interior of  $S$ ). The key to the skein relation is to consider replacing  $M_R$  with some substitutes, as shown in c

last section, the physical Hilbert spaces  $\mathcal{H}_L$  and  $\mathcal{H}_R$  associated with the boundaries of  $M_L$  and  $M_R$  are two dimensional.

The strategy is now the same as the strategy which led to the multiplicativity relation (4.1). The Feynman path integral on  $M_L$  determines a vector  $\chi$  in  $\mathcal{H}_L$ . The Feynman path integral on  $M_R$  determines a vector  $\psi$  in  $\mathcal{H}_R$ . The vector spaces  $\mathcal{H}_L$  and  $\mathcal{H}_R$  (which are associated with the same Riemann surface  $S^2$  with opposite orientation) are canonically dual, and the partition function or Feynman path integral  $Z(L)$  is equal to the natural pairing

$$Z(L) = (\chi, \psi). \quad (4.8)$$

We cannot evaluate (4.8), since we know neither  $\chi$  nor  $\psi$ . The one thing that we do know, at present, is that (for the groups and representations we are considering) this pairing is occurring in a two dimensional vector space. A two dimensional vector space has the marvelous property that any three vectors obey a relation of

linear dependence. Thus, given any two other vectors  $\psi_1$  and  $\psi_2$  in  $\mathcal{H}_R$ , there would be a linear relation

$$\alpha\psi + \beta\psi_1 + \gamma\psi_2 = 0, \quad (4.9)$$

where  $\alpha$ ,  $\beta$ , and  $\gamma$  are complex numbers. Physically, there is a very natural way to get additional vectors in  $\mathcal{H}_R$ . If one replaces  $M_R$  in Fig. 7b by any other three manifold  $X$  with the same boundary (and with suitable strings in  $X$  connecting the marked points on the boundary of  $M_R$ ), then the Feynman path integral on  $X$  gives rise to a new vector in  $\mathcal{H}_R$ . Picking any two convenient three manifolds  $X_1$  and  $X_2$  for this computation gives vectors  $\psi_1$  and  $\psi_2$  that can be used in (4.9). We will consider the case in which  $X_1$  and  $X_2$  are the same manifolds as  $M_R$  but with different “braids” connecting the points on the boundary; this is indicated in Fig. 7c.

Once  $\psi_1$  and  $\psi_2$  are obtained in this way, (4.9) has the obvious consequence that

$$\alpha(\chi, \psi) + \beta(\chi, \psi_1) + \gamma(\chi, \psi_2) = 0. \quad (4.10)$$

The three terms in (4.10) have a “physical” interpretation, evident in Fig. 7c. By gluing  $M_L$  back together with  $M_R$  or one of its substitutes  $X_1$  and  $X_2$ , one gets back the original three manifold  $M$ , but with the original link  $L$  replaced by some new links  $L_1$  and  $L_2$ . Thus, (4.10) amounts to a relation among the link expectation values of interest, namely

$$\alpha Z(L) + \beta Z(L_1) + \gamma Z(L_2) = 0. \quad (4.11)$$

This recursion relation is often drawn as in Fig. 8. The meaning of this figure is as follows. If one considers three links whose plane projections are identical outside a disc, and look inside this disc like the three drawings in the figure, then the expectation values of those links, weighted with coefficients  $\alpha$ ,  $\beta$ , and  $\gamma$ , add to zero.

It is well known in knot theory that (4.11) uniquely determines the expectation values of all knots in  $S^3$ . For convenience we include a brief explanation of this. One starts with a plane projection of a knot, indicated in Fig. 9. The number  $p$  of crossings is finite. Inductively, suppose that all knot expectation values for knots with at most  $p - 1$  crossings have already been computed. One wishes to study knots with  $p$  crossings. If one had  $\beta = 0$  in (4.11), one could at each crossing pass the two strands through each other with a factor of  $-\gamma/\alpha$  in replacing an over-crossing by an under-crossing. If this were possible, the lines would be effectively transparent, and one could untie all knots. As it is,  $\beta \neq 0$ , but the term proportional to  $\beta$  reduces the number of crossings, giving rise to a new link whose expectation value is already known by the induction hypothesis.

This process reduces the discussion to the case  $p = 0$  where there are no crossings, and therefore we are dealing only with a certain number of unlinked and unknotted circles. For practice with (4.11), let us discuss this case explicitly. In Fig. 10, we sketch a useful special case of Fig. 8. The first and third links in links in Fig. 10 consist of a single unknotted circle, and the second consists of two unlinked and unknotted circles. If we denote the partition function for  $s$  unlinked

$$\alpha \left( \text{crossing} \right) + \beta \left( \text{parallel lines} \right) + \gamma \left( \text{crossing} \right) = 0$$

Fig. 8. A recursion relation for links

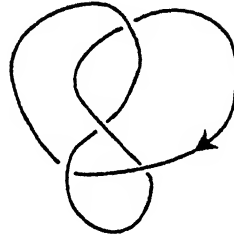


Fig. 9. A plane projection of a knot, with four crossings

$$\alpha \left( \text{circle with crossing} \right) + \beta \left( \text{circle with crossing} \right) + \gamma \left( \text{circle with crossing} \right) = 0$$

Fig. 10. A special case of the use of Fig. 8. The idea is that the three pictures are identical outside of the dotted lines, and look like figure (4.11) inside them

and unknotted circles in the  $N$  dimensional representation of  $SU(N)$  as  $Z(S^3; Cs)$ , then (4.11) amounts in this case to the assertion that

$$(\alpha + \gamma) Z(S^3; C) + \beta Z(S^3; C^2) = 0. \quad (4.12)$$

Together with (4.7),<sup>10</sup> this implies that the expectation value of an unknotted Wilson line in the  $N$  dimensional representation of  $SU(N)$  is

$$\langle C \rangle = -\frac{\alpha + \gamma}{\beta}. \quad (4.13)$$

Presently we will make this formula completely explicit by computing  $\alpha$ ,  $\beta$ , and  $\gamma$  in terms of the fundamental quantum field theory parameters  $N$  and  $k$ .

The induction sketched above expresses any knot expectation value as a rational function of  $\alpha$ ,  $\beta$  and  $\gamma$  (a ratio of polynomials), after finitely many steps. It is in this sense that the Jones knot invariants and their generalizations are “polynomials”. While it is, as we have seen, comparatively elementary to prove that (4.11) uniquely determines the knot invariants, the converse is far less obvious. Equations (4.11) can be used in many different ways to obtain the expectation value of a given link, and one must show that one does not run into

<sup>10</sup> This is the only point at which (4.7) has to be used. The induction sketched in the previous paragraph reduces all computations for knots in  $S^3$  to this special case without using (4.7)



any inconsistency. While this has been proved in a variety of ways, the proofs have not been intrinsically three dimensional – (4.11) has not previously been derived from a manifestly invariant three dimensional framework. This is the novelty of the present discussion.

*Change of Framing.* We want to compute  $\alpha$ ,  $\beta$ , and  $\gamma$ , but as a prelude we must discuss a certain technical point. At the end of Sect. 2, we learned that choosing a circle  $C$  and a representation  $R$  is not enough to give a well defined quantum holonomy operator  $W_R(C) = \text{Tr}_R P \exp \int_C \text{Ad } x$ . It is also necessary to pick a “framing” of the circle  $C$ , which enters when one has to calculate the self-linking number of  $C$  and its non-abelian and quantum generalizations. At the end of Sect. 2, we promised to derive a formula (2.33) showing how any partition function with an insertion of  $W_R(C)$  transforms under a change of framing. Now it is time to deliver on this promise.

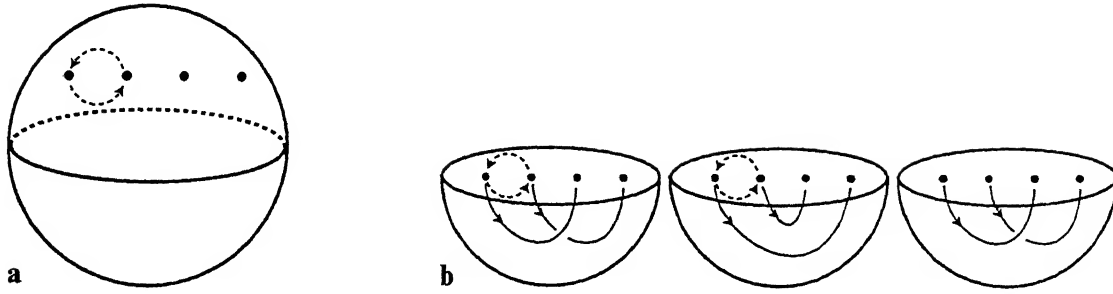
As in Fig. 7b, let us cut the three manifold  $M$  on a Riemann surface  $\Sigma$  that intersects  $C$  in a point  $P$  (and perhaps in some other points that will not be material). In our previous argument, we used the fact that associated with the boundaries  $M_L$  or  $M_R$  are Hilbert spaces  $\mathcal{H}_L$  and  $\mathcal{H}_R$ . Moreover,  $\mathcal{H}_R$  (for example) is “a flat bundle on moduli space” so the mapping class group of the boundary  $\Sigma$  acts naturally on  $\mathcal{H}_R$ . We wish to act on the boundary of  $M_R$  with a very particular diffeomorphism before gluing the pieces of Fig. 7b back together again. The diffeomorphism that we want to pick is a  $t$ -fold “Dehn twist” about the point  $P$  on  $\Sigma$ . Making this diffeomorphism and then gluing the pieces of Fig. 7b back together again, one gets an identical looking picture, but the framing of the circle  $C$  has been shifted by  $t$  units. On the other hand, one knows in conformal field theory how the Dehn twist acts on  $\mathcal{H}_R$ . Associated with the representation  $R$  is a number  $h_R$ , the “conformal weight of the primary field in the  $R$  representation”. The  $t$ -fold Dehn twist acts on  $\mathcal{H}_R$  as multiplication by  $e^{2\pi i t h_R}$ . So we have obtained (2.33) with  $h = h_R$ .

*Explicit Evaluation.* We will now determine the parameters  $\alpha$ ,  $\beta$ , and  $\gamma$  that appear in the crucial equation (4.10). We need to determine the explicit relation among the three vectors  $\psi$ ,  $\psi_1$  and  $\psi_2$  that appear in Fig. 7c. This requires a further study of the two dimensional Hilbert space which arises as the space of conformal blocks for the  $R, R, \bar{R}, \bar{R}$  four point function on  $S^2$  [ $R$  being in this case the defining  $N$  dimensional representation of  $SU(N)$  and  $\bar{R}$  its dual]. The three configurations in Fig. 7c can be regarded as differing from each other by a certain diffeomorphism of  $S^2$ ; the diffeomorphism in question is the “half-monodromy” under which the two copies of  $R$  change places by taking a half-step around one another, as indicated in Fig. 11. Moore and Seiberg call this operation  $B$  and study it extensively. The states  $\psi_1$  and  $\psi_2$  are none other than

$$\psi_1 = B\psi, \quad \psi_2 = B^2\psi. \quad (4.14)$$

The matrix  $B$ , since it acts in a two dimensional space, obeys a characteristic equation

$$B^2 - yB + z = 0, \quad (4.15)$$



**Fig. 11 a and b.** The half-monodromy operation exchanging two equivalent points on  $S^2$  is sketched in **a**; the arrows are meant to suggest a process in which the first two points change places by executing a half-twist about one another. The idea in **b** is that if the two points on the left in the first picture undergo a half-twist about one another, the first picture becomes the second, and if this is done again, the second picture becomes the third. In this way the three pictures on the right of Fig. 7c differ by a succession of half-monodromies

where

$$y = \text{Tr } B, \quad z = \det B. \quad (4.16)$$

In view of (4.14), the linear relation among  $\psi$ ,  $\psi_1$ , and  $\psi_2$  is (up to an irrelevant common factor) just

$$z \cdot \psi - y \cdot \psi_1 + \psi_2 = 0, \quad (4.17)$$

and according to (4.16), to make this explicit we need only to know the eigenvalues (and thus the determinant and trace) of  $B$ .

These can be obtained from [13], but before describing the formulas, I would like to point out an important subtlety. As we have discussed in the last subsection, all concrete results such as the values of  $\alpha$ ,  $\beta$ , and  $\gamma$  depend on the framing of knots. The convention that is most natural in working on an arbitrary three manifold is not the convention usually used in discussing knots on  $S^3$ .

In studying Fig. 7c, to describe the relative framings, the task is to specify the relative framing of the three pictures on the right, since the picture on the left is being held fixed. If one just looks at these three pictures and ignores the fact that the lines cannot pass through each other, there is an obvious sense in which one would like to pick “the same” framing for each picture; for instance, a unit vector coming out of the page defines a normal vector field on each link in the picture.

This is equivalent to the convention of Moore and Seiberg in defining the eigenvalues of  $B$ , so we can now quote their results. Let  $h_R$  be the conformal weight of a primary conformal field transforming as  $R$ , let  $E_i$  be the irreducible representations of  $SU(N)$  appearing in the decomposition of  $R \otimes R$ , and let  $h_{E_i}$  be the weights of the corresponding primary fields. Then the eigenvalues of  $B$  are

$$\lambda_i = \pm \exp(i\pi(2h_R - h_{E_i})), \quad (4.18)$$

where the  $+$  or  $-$  sign corresponds to whether  $E_i$  appears symmetrically or antisymmetrically in  $R \otimes R$ . If  $R$  is the  $N$  dimensional representation of  $SU(N)$ , then one finds<sup>11</sup> that the eigenvalues of  $B$  are

$$\lambda_1 = \exp\left(\frac{i\pi(-N+1)}{N(N+k)}\right), \quad \lambda_2 = -\exp\left(\frac{i\pi(N+1)}{N(N+k)}\right). \quad (4.19)$$

<sup>11</sup> For this representation,  $h_R = (N^2 - 1)/(2N(N+k))$ . In the decomposition of  $R \otimes R$ , the symmetric piece is an irreducible representation with  $h_{E_1} = (N^2 + N - 2)/N(N+k)$ , and the antisymmetric piece is an irreducible representation with  $h_{E_2} = (N^2 - N - 2)/N(N+k)$

It is straightforward to put these formulas in (4.16), (4.17) and thus make our previous results completely explicit.

Before comparing to the knot theory literature, it is necessary to make a correction in these results. For a link in  $S^3$ , there is always a standard framing in which the self-intersection number of each component of the link is zero. Values of the knot polynomials for knots in  $S^3$  are usually quoted without specifying a framing; these are the values for the link with standard framing. However, if on the right of Fig. 7c we use the “same” framing for each picture, then when the right of Fig. 7c is glued to the left, one does not have the canonical framing for each link. If the first knot is framed in the standard fashion, then the second is in error by one unit and the third by two units. So after using (4.19) to compute  $\alpha, \beta, \gamma$ , we must, if we wish to agree with the knot theory literature, multiply  $\beta$  by  $\exp(-2\pi i h_R)$  and  $\gamma$  by  $\exp(-4\pi i h_R)$ . After these corrections, one gets

$$\begin{aligned}\alpha &= -\exp\left(\frac{2\pi i}{N(N+k)}\right), \\ \beta &= -\exp\left(\frac{i\pi(2-N-N^2)}{N(N+k)}\right) + \exp\left(\frac{i\pi(2+N-N^2)}{N(N+k)}\right), \\ \gamma &= \exp\left(\frac{2\pi i(1-N^2)}{N(N+k)}\right).\end{aligned}\tag{4.20}$$

If one multiplies  $\alpha, \beta, \gamma$  by an irrelevant common factor  $\exp(i\pi(N^2-2)/N(N+k))$  and introduces the variable

$$q = \exp(2\pi i/(N+k)),\tag{4.21}$$

then the skein relation can be written more elegantly as

$$-q^{N/2} L_+ + (q^{1/2} - q^{-1/2}) L_0 + q^{-N/2} L_- = 0.\tag{4.22}$$

Here  $L_+, L_0$ , and  $L_-$  [equivalent to  $L, L_1$ , and  $L_2$  in (4.11)] are standard notation for overcrossing, zero crossing, and undercrossing; and for  $i = +, 0, -$ , we now write simply  $L_i$ , instead of  $Z(L_i)$ . Equation (4.22) is correctly normalized to give the right answers for knots on  $S^3$  with their standard framing, and if one is only interested in knots on  $S^3$  one can use it without ever thinking about the framings. Finally, comparing (4.13) and (4.22), we see that the expectation value of an unknotted Wilson line on  $S^3$ , with its standard framing, is

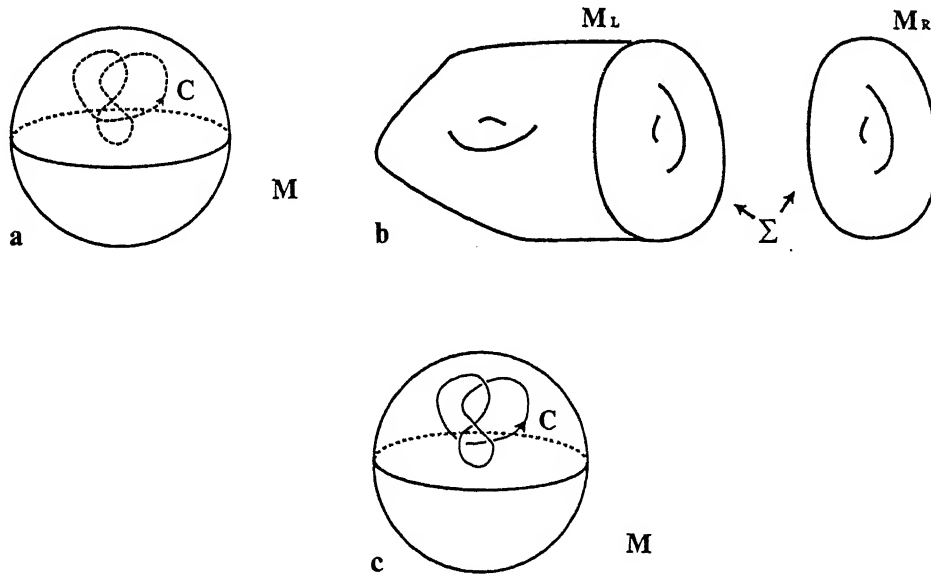
$$\langle C \rangle = \frac{q^{N/2} - q^{-N/2}}{q^{1/2} - q^{-1/2}}.\tag{4.23}$$

This formula can be subjected to several interesting checks. First of all, the right-hand side of (4.23) is positive for all values of the positive integers  $N$  and  $k$ . This is required by reflection positivity of the Chern-Simons gauge theory in three dimensions. Second, in the weak coupling limit of  $k \rightarrow \infty$ , we have  $\langle C \rangle \rightarrow N$ . This is easily interpreted; in the weak coupling limit, the fluctuations in the connection  $A_i$  on  $S^3$  are irrelevant, and the expectation value of the Wilson line approaches its value for  $A_i = 0$ , which is the dimension of the representation, or in this case  $N$ .

#### 4.2. Surgery on Links

We have seen that it is possible to effectively calculate the expectation value of an arbitrary link in  $S^3$ . We would now like to generalize this to computations on an arbitrary three manifold. The basic idea is that by the operation of “surgery on links” any three manifold can be reduced to  $S^3$ , so it is enough to understand how the invariants that we are studying transform under surgery. The operation of surgery can be described as follows. One begins with a three manifold  $M$  and an arbitrarily selected embedded circle  $C$ . Note that there is, to begin with, no Wilson line associated with  $C$ ;  $C$  is simply a mathematical line on which we are going to carry out “surgery”. To do so we first thicken  $C$  to a “tubular neighborhood”, a solid torus centered on  $C$ . Removing this solid torus,  $M$  is split into two pieces; the solid torus is called  $M_R$  in Fig. 12b, and the remainder is called  $M_L$ . One then makes a diffeomorphism on the boundary of  $M_R$  and glues  $M_L$  and  $M_R$  back together to get a new three manifold  $\tilde{M}$ .

It is a not too deep result that every three manifold can be obtained from or reduced to  $S^3$  (or any other desired three manifold) by repeated surgeries on knots. However, such a description is far from unique and it is often difficult to use a description of a three manifold in terms of surgery to compute the invariants of interest. We will now see that the invariants studied in this paper can be effectively computed from a surgery presentation.



**Fig. 12a–c.** Surgery on a circle  $C$  in a three manifold  $M$  is carried out by removing a tubular neighborhood of  $C$ , depicted in a. At this point  $M$  has been separated into two pieces,  $M_L$  and  $M_R$ , with a torus  $\Sigma$  for their boundaries, as sketched in b.  $M_R$  is simply a solid torus. Surgery is completed by gluing the two pieces back together after making a diffeomorphism of the boundary of  $M_R$ . At the end of this process,  $M$  has been replaced by a new three manifold  $\tilde{M}$ . As we will eventually see, computations on  $\tilde{M}$  are equivalent to computations on  $M$  with a physical Wilson line where the surgery was made, as in c. The difference between a and c is that in a the circle  $C$  is just a locale for surgery, but in c it is a Wilson line

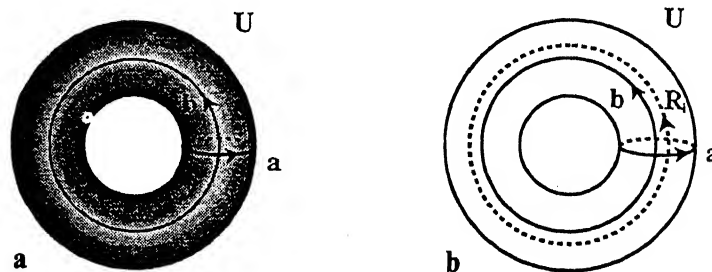
We study Fig. 12b by the standard arguments. Hilbert spaces  $\mathcal{H}_L$  and  $\mathcal{H}_R$ , canonically dual to one another, are associated with the boundaries of  $M_L$  and  $M_R$ . The path integrals on  $M_L$  and  $M_R$  give vectors  $\psi$  and  $\chi$  in  $\mathcal{H}_L$  and  $\mathcal{H}_R$ , and the partition function on  $M$  is just the natural pairing  $(\psi, \chi)$ . If we act on the boundary of  $M_R$  with a diffeomorphism  $K$  before gluing  $M_L$  and  $M_R$  back together, then  $\chi$  is replaced by  $K\chi$  so  $(\psi, \chi)$  is replaced by  $(\psi, K\chi)$ .

This potentially gives a way to determine how the partition function of the quantum field theory transforms under surgery. Upon gaining a suitable understanding of  $K\chi$ , we will be able to reduce calculations on  $\tilde{M}$  to calculations on  $M$ .

#### 4.3. The Physical Hilbert Space in Genus One

At this point we need a description of the physical Hilbert space in genus one. A beautiful description, perfectly adapted for our needs, appears in the work of Verlinde [12].

First of all, the loop group  $LG$  has at level  $k$  finitely many integrable highest weight representations. Let  $t$  be the number of these. For each such highest weight representation of the loop group, the highest weight space is an irreducible representation of the finite dimensional group  $G$ . In this way there appear  $t$  distinguished representations of  $G$ ; we label these as  $R_0, R_1 \dots R_{t-1}$ , with  $R_0$  denoting the trivial representation (which is always one of those on this list). Verlinde showed that if  $\Sigma$  is a Riemann surface of genus one, then the dimension of the physical Hilbert space  $\mathcal{H}_\Sigma$  is  $t$ . Moreover, though there is no canonical basis for  $\mathcal{H}_\Sigma$ , Verlinde showed that every choice of a homology basis for  $H^1(\Sigma, \mathbb{Z})$ , consisting of two cycles  $a$  and  $b$ , gives a canonical choice of basis in  $\mathcal{H}_\Sigma$ . For our purposes, this can be described as follows. Topologically, there are many inequivalent ways to identify a torus  $\Sigma$  as the boundary of a solid torus  $U$ . The choice of  $U$  can be fixed by requiring that the cycle  $a$  is contractible in  $U$ . This is indicated in Fig. 13a. Next, for every  $i = 0 \dots t-1$ , one defines a state  $v_i$  in  $\mathcal{H}_\Sigma$  as follows. One places a Wilson line in the  $R_i$  representation in the interior of  $U$ ,



**Fig. 13a and b.** A Riemann surface  $\Sigma$  of genus one is shown in **a** as the boundary of a solid torus  $U$ ; the indicated  $a$ -cycle is contractible in  $U$ . In **b**, a basis of the physical Hilbert space is indicated consisting of states obtained by placing a Wilson line, in the  $R_i$  representation, in the interior of  $U$ , parallel to the cycle  $b$ , and performing the path integral to get a vector  $v_i$  in  $\mathcal{H}_\Sigma$ .

running in the  $b$  direction,<sup>12</sup> and one performs the Feynman path integral in  $U$  to define a vector  $v_i$  in  $\mathcal{H}_\Sigma$ . The  $v_i$  make up the Verlinde basis in  $\mathcal{H}_\Sigma$ . It must be understood that a Wilson line in the trivial representation is equal to 1, so the vector  $v_0$  obtained by this definition is the same as the vector  $\chi$  which in the last subsection was obtained by a path integral on  $U$  with no Wilson lines:

$$\chi = v_0. \quad (4.24)$$

A diffeomorphism  $K$  of  $\Sigma$  is represented in the Verlinde basis by an explicit matrix  $K_i^j$ , defined by the formula

$$K \cdot v_i = \sum_j K_i^j v_j. \quad (4.25)$$

In the space spanned by the  $v_i$ , there is a natural inner product, defined by the tensor  $g_{ij}$  which is one if  $R_i$  is the dual of  $R_j$  and zero otherwise. We may sometimes use this metric to raise and lower indices, letting  $K_{ij} = \sum_m g_{mj} K_i^m$ .

We can now get a much more concrete description of the behavior of the quantum field theory partition function under surgery. In discussing surgery, we began with a three manifold  $M$  and a knot  $C$ . Cutting out a tubular neighborhood of this knot, whose boundary we call  $\Sigma$ , we separated  $M$  into  $M_L$  and  $M_R$ , with  $M_R$  being a solid torus. The path integrals on  $M_L$  and  $M_R$  gave vectors  $\psi$  and  $\chi$  in the Hilbert spaces  $\mathcal{H}_L$  and  $\mathcal{H}_R$ . As we have just noted,  $\chi$  is the same as  $v_0$ , so the partition function on  $M$  is  $(\psi, v_0)$ . Now we want to make a diffeomorphism  $K$  on the boundary of  $M_R$ , and then glue together  $M_L$  and  $M_R$  to make a new three manifold  $\tilde{M}$ . The partition function of  $\tilde{M}$  is  $Z(\tilde{M}) = (\psi, K v_0)$ , as we saw in the last section. To say that  $Z(\tilde{M})$  is computable from a surgery presentation means that the evaluation of this invariant of  $\tilde{M}$  can be reduced to tractable calculation on  $M$ . We will now show this. From (4.25), we can write

$$Z(\tilde{M}) = \sum_j K_0^j (\psi, v_j). \quad (4.26)$$

But each term  $(\psi, v_j)$  has an interpretation in terms of path integrals on  $M$ ! Indeed, it is the very definition of the  $v_j$  that  $v_j$  differs from  $v_0$  just by an insertion of an extra Wilson line in the  $R_j$  representation at the center of  $M_R$ . So just as  $(\psi, v_0)$  represents the original partition function of  $M$ ,  $(\psi, v_j)$  represents a modified partition function with an extra Wilson line in the  $R_j$  representation placed on  $C$ . So we rewrite (4.26) in the form

$$Z(\tilde{M}) = \sum_j K_0^j \cdot Z(M; R_j), \quad (4.27)$$

where  $Z(M; R_j)$  is the partition function of  $M$  with an extra Wilson line in the  $R_j$  representation included on the circle  $C$  (in addition to whatever Wilson lines are already present on  $M$ ). This is indicated in Fig. 12c. To use (4.27), one needs to know the matrix  $K_i^j$ , which is precisely the matrix by which the diffeomorphism  $K$  of the torus is represented on the characters of the irreducible level  $k$  representations of  $LG$ ; these matrices appear in [37] and have remarkable properties

<sup>12</sup> The  $b$  cycle on the boundary of  $U$  gives a framing of this Wilson line

recently investigated in [12, 13]. Given a knowledge of the  $K_i^j$ , (4.27) is a completely explicit formula expressing computations on  $\tilde{M}$  in terms of computations on  $M$ . By repeated use of this formula, computations on any three manifold can be reduced to computations on  $S^3$ ,<sup>13</sup> with appropriate Wilson lines. Of course, the surgery will generate Wilson lines on  $S^3$  in representations of  $G$  corresponding to arbitrary integrable representations of  $LG$ .

*Generalized Surgery.* The surgery law (4.27) has a useful generalization. While so far we have only considered surgery on a purely imaginary circle  $C$ , as in Fig. 12, there is no reason not to generalize this to a situation in which before the surgery a Wilson line in the  $R_i$  representation was already present on  $C$ . Surgery amounts to cutting out a neighborhood of  $C$  and then gluing it back in, and after this process the  $R_i$  Wilson line will still be present in  $\tilde{M}$ . So the left-hand side of (4.27) is replaced by  $Z(\tilde{M}; R_i)$ , where the notation schematically indicates the presence of the  $R_i$  Wilson line. What about the right-hand side of (4.27)? Before surgery, with a Wilson line  $R_i$  on  $C$ , the path integral on a tubular neighborhood  $U$  of  $C$  gives on the boundary a state  $\chi' = v_i$ ; this is the generalization of (4.24). If we cut out  $C$  and glue it back in with a diffeomorphism  $K$  of the boundary, then  $v_i$  is replaced according to (4.25) with  $K_i^j v_j$ . If we remember that  $v_j$  could have been obtained by putting a Wilson line on  $C$  in the  $R_j$  representation, we see that the right-hand side of (4.27) becomes  $\sum_j K_i^j Z(M; R_j)$ , so we get the generalized surgery formula

$$Z(\tilde{M}; R_i) = \sum_j K_i^j Z(M; R_j). \quad (4.28)$$

This formula will be used later in a new proof of Verlinde's conjecture.

*Cabling of Knots; Satellites.* Finally, let us note that similar methods can be used to determine the behavior of the knot invariants under "cabling", and more generally to relate the invariants of the "satellites" of a knot to invariants of the original knot. Any knot  $C$  in any three manifold  $M$  has a neighborhood that looks like a solid torus  $U$ . If we replace  $C$  by an arbitrary satellite  $\tilde{C}$  of itself (an arbitrary knot that can be placed in  $U$ ), with representations  $R_\sigma$  associated with the connected components of  $\tilde{C}$ , then the path integral on  $U$  will define a vector  $\psi$  in the physical Hilbert space  $\mathcal{H}_\Sigma$  associated with the boundary  $\Sigma$  of  $U$ . Like any vector in  $\mathcal{H}_\Sigma$ ,  $\psi$  can be expanded in the Verlinde basis,

$$\psi = \sum_i \alpha_i v_i. \quad (4.29)$$

The  $\alpha_i$  are complex numbers that depend on the choice of satellite  $\tilde{C}$  and on the choice of representations  $R_\sigma$ , but they do not depend on what three manifold  $M$  the solid torus  $U$  has been extracted from or on what other knots are present on  $M$ . The vectors  $v_i \in \mathcal{H}_\Sigma$  are the vectors that would be produced by the path integral on  $U$  with a Wilson line in the  $R_i$  representation placed on the original knot  $C$  (and not a satellite of  $C$ ). Thus, a knowledge of the invariants of  $C$  in arbitrary representations together with a knowledge of the universal coefficients  $\alpha_i$  is enough to determine the invariants of arbitrary satellites of  $C$ .

<sup>13</sup> Or on any desired three manifold; we will see that  $S^2 \times S^1$  is more tractable than  $S^3$

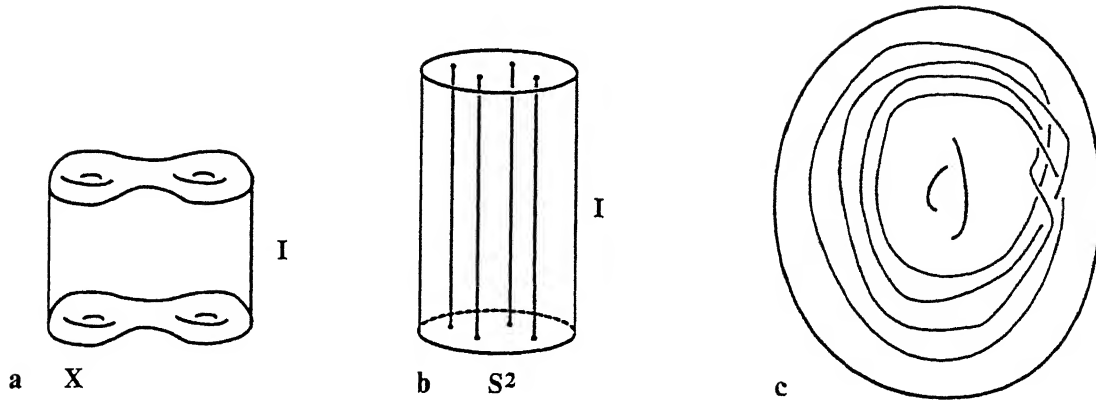


Fig. 14a–c. Beginning with  $X \times I$ , shown in a, one makes  $X \times S^1$  by identifying  $X \times \{0\}$  with  $X \times \{1\}$ . If  $X$  is  $S^2$  with some marked points  $P_i$ , then this construction gives the picture of b. If  $S^2 \times \{0\}$  is joined to  $S^2 \times \{1\}$  via a non-trivial diffeomorphism  $B$ , one makes in this way a braid, as in c

#### 4.4. Path Integrals on $S^1 \times X$

In this subsection we will describe a few facts which are useful in their own right and will enable us to carry out some concrete surgeries.

First of all, we have not so far determined the partition function of  $S^3$  without Wilson lines. It may come as a surprise to topologists that we cannot trivially assert that this is 1. In quantum field theory there is no particularly strong axiom governing the partition function of  $S^3$ . The three manifolds whose partition functions can be computed in a particularly simple way, from the axioms of quantum field theory, are those of the form  $X \times S^1$ , for various  $X$ .  $X \times S^1$  can conveniently be studied in a “Hamiltonian” formalism, as indicated in Fig. 14a. One constructs the Hilbert space  $\mathcal{H}_X$  of  $X$ . Then one introduces a “time” direction, represented by a unit interval  $I = [0, 1]$ , and one propagates the vectors in  $\mathcal{H}_X$  from “time” 0 to “time” 1. This operation is trivial, since the Chern-Simons theory, like any generally covariant theory, has a vanishing Hamiltonian. Finally, one forms  $X \times S^1$  by gluing  $X \times \{0\}$  to  $X \times \{1\}$ ; this identifies the initial and final states, giving a trace:

$$Z(X \times S^1) = \text{Tr}_{\mathcal{H}_X}(1) = \dim \mathcal{H}_X. \quad (4.30)$$

For example, the physical Hilbert space of  $S^2$  is one dimensional, for any  $G$  and  $k$ , so one has

$$Z(S^2 \times S^1) = 1. \quad (4.31)$$

It is possible to generalize (4.30) as follows. If we are given a diffeomorphism  $K: X \rightarrow X$ , then one can form the mapping cylinder  $X \times_K S^1$  by identifying  $x \times \{1\}$  with  $K(x) \times \{0\}$  for every  $x \in X$ . At the level of quantum field theory, when one goes from  $X \times I$  to  $X \times_K S^1$ , the initial and final states are identified via  $K$ , so the generalization of (4.30) is

$$Z(X \times_K S^1) = \text{Tr}_{\mathcal{H}_X}(K). \quad (4.32)$$

The situation that we actually wish to apply this to is the case in which  $X$  is  $S^2$  with some marked points  $P_a$ ,  $a = 1 \dots s$  to which representations  $R_{i(a)}$  are assigned.



[For  $a = 1 \dots s$ ,  $i(a)$  is one of the values  $0 \dots t - 1$  corresponding to integrable level  $k$  representations of the loop group.] In this case, the simple product  $X \times S^1$  is just  $S^2 \times S^1$  with some Wilson lines which are unknotted, parallel circles of the form  $\{P_a\} \times S^1$ , as sketched in Fig. 14b. To determine the path integral on  $S^2 \times S^1$  in the presence of these Wilson lines, which we will denote as  $Z(S^2 \times S^1; \langle R \rangle)$ , one needs to study the Hilbert space of  $S^2$  with charges in the representations  $R_{a_i}$ ; we will denote this as  $\mathcal{H}_{S^2; \langle R \rangle}$ . The analog of (4.31) is then

$$Z(S^2 \times S^1; \langle R \rangle) = \dim \mathcal{H}_{S^2; \langle R \rangle}. \quad (4.33)$$

The dimensions of these spaces were discussed at the end of Sect. 3. Thus, if the collection of representations  $\langle R \rangle$  consists of a single representation  $R_a$ , we get

$$Z(S^2 \times S^1; R_a) = \delta_{a,0}, \quad (4.34)$$

since the physical Hilbert space with a single charge in the  $R_a$  representation is one dimensional if  $R_a$  is the trivial representation ( $a=0$ ) and zero dimensional otherwise.<sup>14</sup> For two charges in the representations  $R_a$  and  $R_b$ , we get

$$Z(S^2 \times S^1; R_a, R_b) = g_{ab}, \quad (4.35)$$

where  $g_{ab}$ , introduced earlier, is 1 if  $R_b$  is the dual of  $R_a$  and zero otherwise. The formula (4.35) follows from the result of Sect. (4.4) for the Hilbert space on  $S^2$  with two charges. Finally, if there are three charges in the representations  $R_a, R_b, R_c$ , we get

$$Z(S^2 \times S^1; R_a, R_b, R_c) = N_{abc}, \quad (4.36)$$

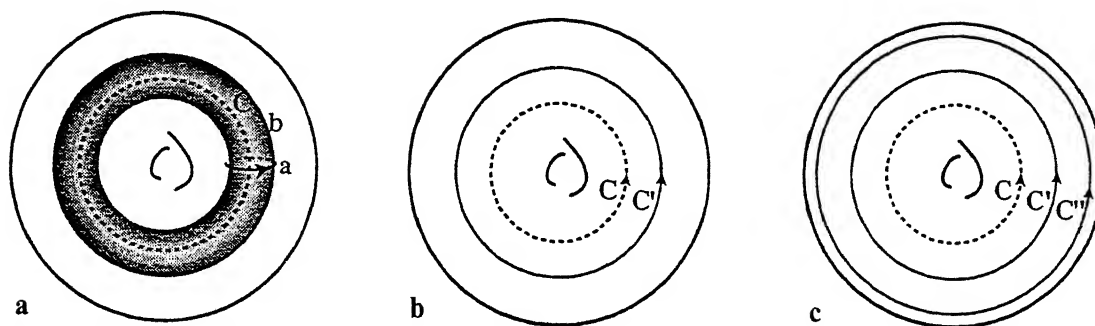
with  $N_{abc}$  the trilinear ‘‘coupling’’ of Verlinde, since this is the dimension of the physical Hilbert space.

#### 4.5. Some Concrete Surgeries

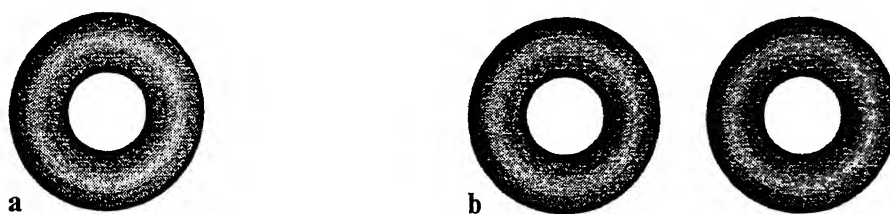
Now we would like to describe some useful results that can be obtained from concrete surgeries. The first goal is compute the partition function of  $S^3$ . Since we already know the partition function of  $S^2 \times S^1$ , we will try to interpret  $S^3$  as a manifold obtained by surgery on  $S^2 \times S^1$ . This is readily done. We consider the circle  $C$  in  $S^2 \times S^1$  indicated in Fig. 15a. A tubular neighborhood of  $C$  is a torus  $\Sigma$ ; we pick a basis of  $H^1(\Sigma; \mathbf{Z})$  consisting of cycles  $a$  and  $b$  indicated in the figure. Now we wish to make a particular surgery associated with a very special diffeomorphism  $S: \Sigma \rightarrow \Sigma$ . We pick  $S$  to map  $a$  to  $b$  and  $b$  to  $-a$ .<sup>15</sup> This surgery – removing the interior of  $\Sigma$  from  $S^2 \times S^1$  and gluing it back after acting with  $S$  – produces a three manifold that is none other than  $S^3$  (Fig. 16). Since this point is crucial in what follows, we pause to explain it. We regard  $S^3$  as  $R^3$  plus a point at infinity. In Fig. 16a a torus  $\Sigma$  has been embedded in  $R^3$ . Obviously,  $\Sigma$  with its

<sup>14</sup> Implicit in (4.34) and subsequent formulas is the use of the standard framing of the Wilson line which is invariant under rotations of  $S^1$ ; it is for this choice that the path integral on  $S^2 \times S^1$  computes the trace of the identity operator in the physical Hilbert space

<sup>15</sup> This transformation, which acts on the upper half plane as  $\tau \rightarrow -1/\tau$ , is indeed usually called  $S$  in the theory of the modular group  $SL(2, \mathbf{Z})$  (which can be identified, via the basis  $a, b$ , with the mapping class group of  $\Sigma$ )



**Fig. 15a–c.** In part a, we consider surgery on a circle  $C$  in  $S^2 \times S^1$ . A tubular neighborhood of this circle is a torus  $\Sigma$ ; a useful basis of  $H^1(\Sigma, \mathbb{Z})$  is indicated. In b, in addition to the circle  $C$  on which we perform surgery, there is a parallel circle  $C'$  on which we place a Wilson line in the  $R_j$  representation. In c, there are two parallel circles  $C'$  and  $C''$  with Wilson lines in representations  $R_j$  and  $R_k$



**Fig. 16a and b.** The purpose of this figure is to indicate how  $S^3$  can be made by surgery starting with  $S^2 \times S^1$ . In a we show a torus  $\Sigma$ , sitting in  $R^3$  and in b a pair of identical solid tori

interior make up a solid torus  $T$ . It is also relatively easy to see that the figure (a) is invariant under inversion, so that the exterior of  $\Sigma$  (including the point at infinity) is a second solid torus  $T'$ . Thus,  $S^3$  can be made by gluing two solid tori along their boundaries. Now in (b) we sketch two identical solid tori  $T$  and  $T'$ ;  $T'$  has been obtained by simply translating  $T$  in Euclidean space. If one glues together the boundaries pointwise with the identification that is indicated by saying that “ $T'$  is a translate of  $T$ ”, one gets  $S^2 \times S^1$ . (In fact, the solid torus  $T$  is  $D \times S^1$ , with  $D$  a two dimensional disc, and  $T'$  is  $D' \times S^1$  with  $D'$  a second disc. Just as two discs  $D$  and  $D'$  glued on their boundary make  $S^2$ ,  $D \times S^1$  naturally glues to  $D' \times S^1$  to make  $S^2 \times S^1$ .) On the other hand, we know from part (a) that  $S^3$  can be obtained by gluing two solid tori. A little mental gymnastics, comparing the argument we gave in connection with (a) to that in (b), shows that to make  $S^3$  we must glue together  $T$  and  $T'$  after making the modular transformation  $S$  on the boundary of  $T'$ .

Now we can use (4.27), with  $S^3$  playing the role of  $\tilde{M}$ ,  $S^2 \times S^1$  playing the role of  $M$ , and the arbitrary diffeomorphism  $K$  replaced by  $S$ . So we learn

$$Z(S^3) = \sum_j S_0^j Z(S^2 \times S^1; R_j). \quad (4.37)$$

We have learned in the last section that  $Z(S^2 \times S^1; R_j)$  is 1 for  $j=0$  and 0 otherwise, so

$$Z(S^3) = S_{0,0}. \quad (4.38)$$

Here  $S_{0,0}$  can be determined from the theory of affine Lie algebras; for  $G = SU(2)$  one gets the formula stated earlier in (2.26). In fact, the whole matrix  $S_{ij}$  can be written very explicitly for  $G = SU(2)$ . The integrable representations of level  $k$  are those of spin  $n/2$  for  $n = 0 \dots k$ , and the matrix elements of  $S$  are

$$S_{mn} = \sqrt{\frac{2}{k+2}} \sin \left( \frac{(m+1)(n+1)\pi}{k+2} \right). \quad (4.39)$$

*The Phase of the Partition Function.* Now let us re-examine in the light of these methods a thorny question that appeared in Sect. 2 – the framing of three manifolds, and the phase of the partition function.

We have obtained  $S^3$  from  $S^2 \times S^1$ , by performing surgery on a certain circle  $C$ , using the modular transformation  $S$ :  $\tau \rightarrow -1/\tau$ . Apart from  $S$ , there are other modular transformations that could be used to build  $S^3$  by surgery on the same knot  $C$  in  $S^2 \times S^1$ . The general choice would be  $T^n S T^m$ , with  $n$  and  $m$  being arbitrary integers, and  $T$  being the modular transformation  $T$ :  $\tau \rightarrow \tau + 1$ . [ $S$  and  $T$  are the standard generators of the modular group, obeying  $S^2 = (ST)^3 = 1$ .] Had we used  $T^n S T^m$ , we would have gotten not (4.38) but

$$Z(S^3) = (T^n S T^m)_{0,0}. \quad (4.40)$$

This may readily be evaluated. In the Verlinde basis,  $T$  is a diagonal matrix with  $T \cdot v_i = e^{2\pi i(h_i - c/24)} \cdot v_i$ ;  $h_i$  is the conformal weight of the primary field in the representation  $R_i$  and  $c$  is the central charge for current algebra with symmetry group  $G$  at level  $k$ . Since  $h_0 = 0$ , if we replace (4.38) by (4.40) the partition function transforms as

$$Z \rightarrow Z \cdot \exp \left( 2\pi i(n-m) \cdot \frac{c}{24} \right). \quad (4.41)$$

Though we have obtained this formula in the example of a particular surgery (giving  $S^3$  from  $S^2 \times S^1$ ), the same ambiguity arises in any process of surgery. Whenever one makes surgery on a circle  $C$ , in a three manifold  $M$ , with the surgery being determined by an  $SL(2, \mathbb{Z})$  element  $u$ , one could instead consider surgery on the same circle  $C$ , using the  $SL(2, \mathbb{Z})$  element  $u \cdot T^m$ . This would have the same effect topologically, but our surgery law would give a partition function containing an extra phase  $\exp(-2\pi i m \cdot c/24)$ .

This phase ambiguity was already encountered, in the large  $k$  limit, in formula (2.24). What is more, from the discussion in Sect. 2, we know what topological structure on three manifolds must be considered in order to keep track of the factors of  $\exp(2\pi i \cdot c/24)$ . One must consider “framed” three manifolds. Two surgeries that have the same effect on the topology of a three manifold may have different effects on the framing. I will discuss elsewhere how to systematically keep track of the factors of  $\exp(2\pi i \cdot c/24)$  under surgery. In the simple applications in this paper, this will not be necessary. All of our applications will involve considering the standard surgery (by the modular transformation  $S$ ) that was used in the last subsection to obtain  $S^3$  from  $S^2 \times S^1$ .

*Some Expectation Values.* Now let us see if we can go farther and determine the path integral  $Z(S^3; R_j)$  on  $S^3$  with an unknotted Wilson line on  $S^3$  in an arbitrary

representation  $R_j$ .<sup>16</sup> To do this, we start on  $S^2 \times S^1$  with a Wilson line in the  $R_j$  representation running parallel to the circle  $C$  on which we are doing surgery, as in Fig. 15b. Carrying out the same surgery as before turns  $S^2 \times S^1$  into  $S^3$ , with a Wilson line in the  $R_j$  representation on  $S^3$ . Application of (4.27) now gives

$$Z(S^3; R_j) = \sum_i S_0^i Z(S^2 \times S^1; R_i, R_j). \quad (4.42)$$

Using (4.35), we can evaluate this and determine the partition function for a Wilson line in an arbitrary representation  $R_j$ ; it is

$$Z(S^3; R_j) = \sum_i S_0^i g_{ij} = S_{0,j}. \quad (4.43)$$

Let us compare this to our previous evaluation (4.23) of the expectation value of an unknotted Wilson line in  $S^3$ . We must recall that the symbol  $\langle C \rangle$  in (4.23) represented a ratio  $\langle C \rangle = Z(S^3; R)/Z(S^3)$ . Let us take  $G = SU(2)$ , so that we can use the explicit formulas (4.39), and take  $R$  to be the two dimensional representation of  $SU(2)$ , so that we can compare to (4.23). Using (4.38), (4.43), and (4.39), we get

$$\langle C \rangle = \frac{S_{0,1}}{S_{0,0}} = \frac{\sin(2\pi/(k+2))}{\sin(\pi/(k+2))}. \quad (4.44)$$

It is easy to see that setting  $N = 2$  in (4.23) gives the same formula. Let us take this one step further and try to calculate by these methods the partition function  $Z(S^3; R_j, R_k)$  for  $S^3$  with two unknotted, unlinked Wilson lines in representations  $R_j$  and  $R_k$ . In Fig. 15c, we start on  $S^2 \times S^1$  with *two* Wilson lines, in representations  $R_j$  and  $R_k$ , parallel to the circle  $C$  on which surgery is to be performed. Carrying out the surgery, we get to  $S^3$  with the desired unlinked, unknotted circles. In this case, the surgery formula (4.27) tells us that

$$Z(S^3; R_j, R_k) = \sum_i S_0^i Z(S^2 \times S^1; R_i, R_j, R_k). \quad (4.45)$$

The right-hand side can be evaluated with (4.36), while the left-hand side can be reduced to (4.43) using (4.6). We get

$$\frac{S_{0,j} S_{0,k}}{S_{0,0}} = \sum_i S_0^i N_{ijk}. \quad (4.46)$$

*Proof of Verlinde's Conjecture.* The last equation is a special case of a celebrated conjecture by Verlinde, which has been proved by Moore and Seiberg [13]. We can use these methods to give a new proof of Verlinde's conjecture, in the case of current algebra. We will have to use the generalized surgery relation (4.28). We return to Fig. 15b but now instead of treating  $C$  as a purely imaginary contour on which surgery is to be performed, we suppose that there is a Wilson line on  $C$  in the  $R_i$  representation. In this case, the standard surgery on  $C$  will still turn  $S^2 \times S^1$  into  $S^3$ , but now on  $S^3$  we will have two Wilson lines, in the  $R_i$  and  $R_j$  representations. Some mental gymnastics shows that they are linked, as in Fig. 17a; schematically,

<sup>16</sup> We give this Wilson line the framing described in the footnote after (4.34); after surgery this turns into the standard framing on  $S^3$

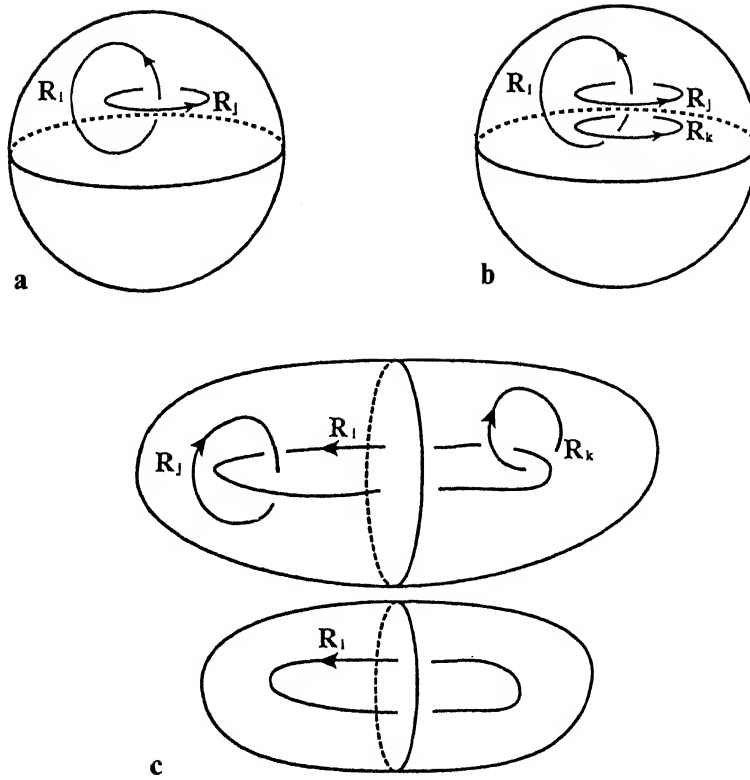


Fig. 17a–c. In a, we sketch linked but unknotted Wilson lines on  $S^3$ , in the  $R_i$  and  $R_j$  representations. In b, a Wilson line  $R_i$  is linked with two Wilson lines  $R_j$  and  $R_k$  on  $S^3$ . In c, we sketch how two crucial amplitudes can be factored through the same one dimensional space

we refer to this linked pair of Wilson lines as  $L(R_i; R_j)$ . The use of (4.28) therefore determines the partition function of  $S^3$  with a pair of linked Wilson lines:

$$Z(S^3; L(R_i; R_j)) = \sum_k S_i^k Z(S^2 \times S^1; R_k, R_j) = S_{ij}. \quad (4.47)$$

In the second step, we have used the fact that the partition function of  $S^2 \times S^1$  with static charges  $R_k$  and  $R_j$  is the metric that we have called  $g_{kj}$  – one if  $R_k$  is dual to  $R_j$  and otherwise zero.

Now let us go back to Fig. 15c, and again on what was previously the purely imaginary circle  $C$  we put a Wilson line in the  $R_i$  representation. The standard surgery on this link will now produce a picture sketched in Fig. 17b, with a Wilson line on  $S^3$  in the  $R_i$  representation that links a pair of Wilson lines  $R_j, R_k$  that are themselves unlinked and unknotted. We call this configuration  $L(R_i; R_j, R_k)$ . The evaluation of (4.28) now gives

$$Z(S^3; L(R_i; R_j, R_k)) = \sum_m S_i^m Z(S^2 \times S^1; R_m, R_j, R_k) = \sum_m S_i^m N_{mjk}. \quad (4.48)$$

To obtain Verlinde's formula, it is now necessary to find an independent way to evaluate the left-hand side.

Such a method is provided by the following generalization of the multiplicativity formula (4.1). The key point in the derivation of (4.1) was that the physical Hilbert space for  $S^2$  with no charges was one dimensional. It is likewise true that the physical Hilbert space  $\mathcal{H}$  for  $S^2$  with a pair of charges in the dual

representations  $R_i$  and  $R_{\bar{i}}$  is one dimensional. Using this and otherwise repeating the derivation of (4.1) gives the following formula:

$$Z(S^3; L(R_i; R_j, R_k)) \cdot Z(S^3; R_i) = Z(S^3; L(R_i; R_j)) \cdot Z(S^3; L(R_i; R_k)). \quad (4.49)$$

The idea in (4.49) is that, as in Fig. 17c, the evaluation of  $Z(S^3; L(R_i; R_j, R_k))$  can be expressed as a pairing  $(\psi, \chi)$  where  $\psi$  and  $\chi$  are certain vectors in  $\mathcal{H}$  and its dual. Likewise the evaluation of  $Z(S^3; R_i)$  is a pairing  $(v', v)$ , where  $v'$  and  $v$  are vectors in  $\mathcal{H}$  and its dual. Using the wonderful fact of one dimensional linear algebra  $(\psi, \chi) \cdot (v', v) = (\psi, v) \cdot (v', \chi)$ , we arrive at (4.49). Since all factors in (4.49) are known except the first, we arrive at the result

$$Z(S^3; L(R_i; R_j, R_k)) = S_{ij} S_{ik} / S_{0,i}. \quad (4.50)$$

Combining this with (4.48), we have

$$S_{ij} S_{ik} / S_{0,i} = \sum_m S_i^m N_{mjk}. \quad (4.51)$$

This is equivalent to Verlinde's statement that "the matrix  $S$  diagonalizes the fusion rules". In other words, in the basis  $v_i$  indicated in Fig. 13, the structure constants of the Verlinde algebra are by definition  $v_i v_j = \sum_k N_{ij}^k v_k$ , where  $N_{ij}^k = \sum_r N_{ijr} g^{rk}$ . If we introduce a new basis  $w_i = S_{0,i} \cdot \sum_m S_i^m v_m$ , then the Verlinde algebra reduces to  $w_i w_j = \delta_{ij} w_j$ . To verify this, we compute

$$w_i w_j = \sum_{k,l} S_i^k S_j^l v_k v_l \cdot S_{0,i} S_{0,j}. \quad (4.52)$$

Using  $v_k v_l = \sum_m N_{kl}^m v_m$  and (4.51), this becomes

$$w_i w_j = S_j^l v^m \cdot S_{il} S_i^m \cdot S_{0,j}. \quad (4.53)$$

Using the unitarity of  $S$ , in the form  $S_j^l S_{il} = \delta_{ij}$ , we see that

$$w_i w_j = \delta_{ij} \sum_m S_j^m v_m \cdot S_{0,j} = \delta_{ij} \cdot w_j, \quad (4.54)$$

showing that the Verlinde algebra has been diagonalized and that the  $w_i$  are idempotents.

*Connected Sum of Links.* At the beginning of Sect. 4, we have seen that the quantum partition function has a multiplicative behavior under connected sum of three manifolds. From (4.1), if  $M = M_1 + M_2$ , then  $Z(M) \cdot Z(S^3) = Z(M_1) \cdot Z(M_2)$ . In the special case that  $M_1$  and  $M_2$  are copies of  $S^3$  with links in them, the connected sum of  $M_1$  and  $M_2$  is a copy of  $S^3$  containing the *disconnected* sum of the two links.

In knot theory there is also an operation of taking the *connected* sum of two links. This operation has appeared in the above discussion. The link that we have called  $L(R_i; R_j, R_k)$  is the connected sum of the two links that we have called  $L(R_i, R_j)$  and  $L(R_i, R_k)$ . In fact, in Fig. 17c,  $L(R_i; R_j, R_k)$  is "cut" into two pieces, which are respectively  $L(R_i, R_j)$  and  $L(R_i, R_k)$  with in each case a connected segment removed. This is the defining configuration for the "connected sum of links". Accordingly, the reasoning that led to (4.49) has the following more

general consequence. If  $L_1$  and  $L_2$  are two links, and  $L_1 + L_2$  is their connected sum, then

$$Z(S^3; L_1 + L_2) \cdot Z(S^3; C) = Z(S^3; L_1) \cdot Z(S^3; L_2). \quad (4.55)$$

Here it is understood that [as in (4.49)] representations have been assigned to the connected components of  $L_1$ ,  $L_2$ , and  $L_1 + L_2$  in a compatible fashion;  $C$  is an unknot placed in whatever representation is carried by the strand “cut” in the generalization of Fig. 17c. Equation (4.55) has a generalization in which  $L_1$  and  $L_2$  are links in arbitrary three manifolds  $M_1$  and  $M_2$ ; then the connected sum of links  $L_1 + L_2$  is a link in the connected sum of manifolds  $M = M_1 + M_2$ , and (4.55) is replaced by

$$Z(M_1 + M_2; L_1 + L_2) \cdot Z(S^3; C) = Z(M_1; L_1) \cdot Z(M_2; L_2). \quad (4.56)$$

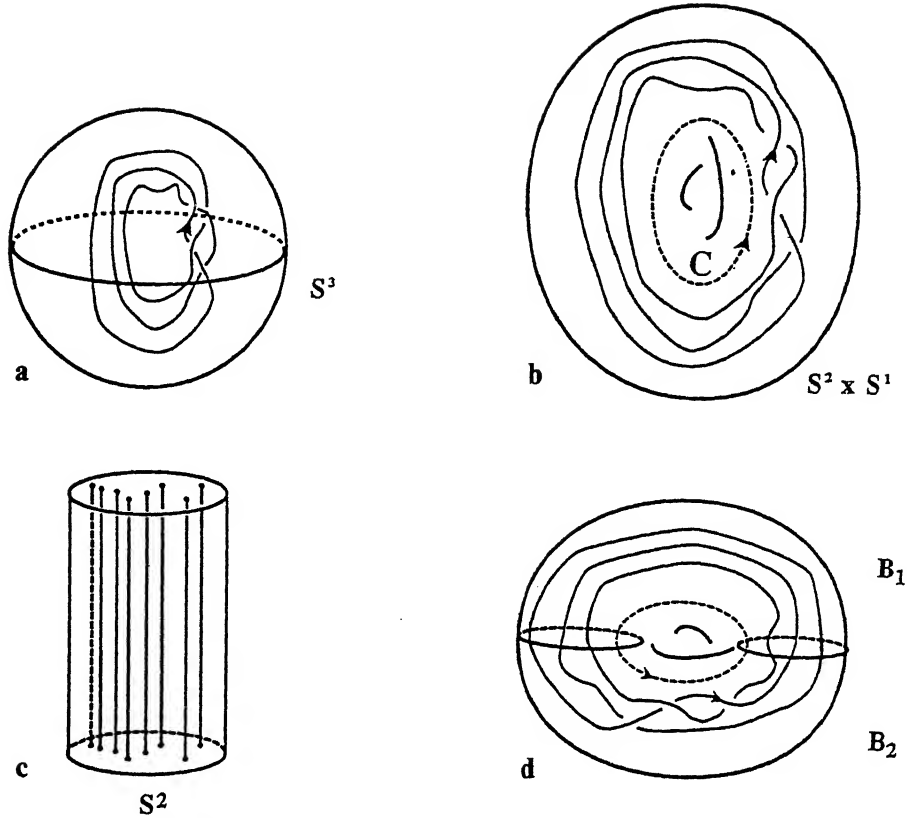
#### 4.6. The Knot Polynomials and the Braid Group

The results in the last subsection are nice enough so that one may wonder if the partition function for an arbitrary link on  $S^3$  can be evaluated in this way. This can indeed be done, and in a way that is closely related to the original route by which the Jones polynomial was discovered, though we cannot expect such explicit formulas as in the simple cases treated above. An arbitrary link  $L$  on  $S^3$ , whose partition function we will call  $Z(S^3; L)$ , can be arranged in the form of a braid, as indicated in Fig. 18a. One can imagine “lifting” this braid  $B$  out of  $S^3$  and putting it on  $S^2 \times S^1$ . To get back to  $S^3$  one would have to do surgery on a circle running parallel to the braid, as suggested in Fig. 18b. The general surgery formula (4.27) then tells us

$$Z(S^3; L) = \sum_j S_0^j Z(S^2 \times S^1; R_j, B), \quad (4.57)$$

where  $Z(S^2 \times S^1; R_j, B)$  is the partition function on  $S^2 \times S^1$  in the presence of both the braid  $B$  and a parallel Wilson line in the  $R_j$  representation. We want to rewrite this in the spirit of (4.32). Suppose that the braid  $B$  contains  $n$  strands making up a collection of representations  $\langle R \rangle$ . Then  $B$  can be regarded as defining an element of the Artin braid group on  $n$  letters. The braid group is closely related to the mapping class group for  $S^2$  with marked points. The reason for that is that if in Fig. 18b we “cut”  $S^2 \times S^1$ , to get back to  $S^2 \times I$  (this amounts to undoing what was done in Fig. 14, then the braid can be unbraided. Thus, the complete information about the braid is in the choice of a diffeomorphism of  $S^2$  (constrained to preserve the marking of the points) by which the top and bottom of Fig. 18c are to be identified. This, however, does not quite mean that the braid group is the same as the mapping class group. In Fig. 18b, there are  $n + 1$  strands, one of which arose from the surgery and does not participate in the braid, while the other  $n$  strands make up the braid. The braid group on  $n$  letters is the subgroup of the mapping class group on  $n + 1$  letters which fixes one of the (framed) strands. There are a number of invariant traces on the braid group that can be naturally defined with the data at our disposal. They are

$$\tau_i(B) = Z(S^2 \times S^1; R_i, B). \quad (4.58)$$



**Fig. 18a–d.** Any link on  $S^3$  can be shaped into the form of a braid, as in **a**; putting the same braid on  $S^2 \times S^1$ , and doing surgery on the circle  $C$  indicated with the dotted line in part **b**, one gets back to the original link on  $S^3$ . If one “cuts”  $S^2 \times S^1$  to get to  $S^2 \times I$ , the braid can be “unbraided”; the braid is recovered by prescribing a diffeomorphism of  $S^2$ , via which the top and bottom of part **c** are to be identified. In **d**, we sketch the origin of the key property of the braid traces. In the presence of an arbitrary Wilson line on the dotted contour (reflecting the results of surgery), two braids  $B_1$  and  $B_2$  are joined end to end in  $S^2 \times S^1$ . There is no way to tell which comes first, so the partition function is invariant under exchange of  $B_1$  and  $B_2$ .

This has the key property of a trace,

$$\tau_i(B_1 B_2) = \tau_i(B_2 B_1), \quad (4.59)$$

since the two sides of (4.59) have the same path integral representation, in which the two braids are glued end to end in  $S^2 \times S^1$ , as in Fig. 18d. Not only does (4.58) obey (4.59); it is actually equal to the trace of the operator  $B$  in a certain representation of the braid group, namely the representation furnished by the physical Hilbert space  $\mathcal{H}$  for  $S^2$  with the  $n+1$  charges, it being understood that the braid group is acting on the first  $n$  charges, and the  $(n+1)^{\text{st}}$  is fixed in the  $R_i$  representation. That  $\tau_i(B)$  is the trace of  $B$  in this Hilbert space is a statement just along the lines of (4.32).

So we can rewrite (4.57) in the form

$$Z(S^3; L) = \sum_j S_0^j \tau_j(B). \quad (4.60)$$

This shows that the link invariants on  $S^3$  may be written as linear combinations of braid traces. This is very close to how the knot polynomials were originally



discovered. It is clear from the work of Tsuchiya and Kanie [11] and Segal [16] along with what has been said above that the braid traces that arise from the Chern-Simons theory are precisely those that first appeared in the work of Jones.

## 5. Applications To Physics

Finally, I would like to comment on the likely implications of these results for physics. We have been exploring a three dimensional viewpoint about conformal field theory, at least for the important special case of current algebra on Riemann surfaces. Many aspects of rational conformal field theory have emerged as natural consequences of general covariance in three dimensions. It seems likely that the marvelous hexagons and pentagons of [13], and the other consistency conditions of rational conformal field theories, can be synthesized by saying that such theories come from generally covariant theories in three dimensions. If so, general covariance in three dimensions may well emerge as one of the main unifying themes governing two dimensional conformal field theory. Such considerations have motivated a study of  $2 + 1$  dimensional gravity which will appear elsewhere [38].

The basic connection that we have so far stated between general covariance in  $2 + 1$  dimensions and conformally invariant theories in  $1 + 1$  dimensions is that the physical Hilbert spaces obtained by quantization in  $2 + 1$  dimensions can be interpreted as the spaces of conformal blocks in  $1 + 1$  dimensions. This connection may seem rather abstract, and I will now make a few remarks aimed at making a more concrete connection. Starting from three dimensions we were led to the problem of quantizing the Lagrangian (3.1), which we repeat for convenience:

$$\mathcal{L} = \frac{k}{8\pi} \int dt \int_{\Sigma} \varepsilon^{ij} \text{Tr} A_i \frac{d}{dt} A_j. \quad (5.1)$$

This is to be quantized with constraints  $\varepsilon^{ij} F_{ij}^a = 0$ , these constraints being the generators of the infinitesimal gauge transformations

$$A_i \rightarrow A_i - D_i \varepsilon. \quad (5.2)$$

So far we have quantized (5.1) on Riemann surfaces without boundary, but now let us relax this requirement. Quantization of (5.2) in this more general case amounts to studying the three dimensional Chern-Simons theory on a three manifold with boundary, namely  $\Sigma \times R^1$ , where  $\Sigma$  is a Riemann surface with boundary. For instance, let  $\Sigma$  be a disc  $D$ . To quantize (5.1) on the disc, we must impose the constraints, which generate the gauge transformations (5.2). As  $D$  has a boundary, we must choose boundary conditions on  $A_i$  and  $\varepsilon$  (for closed surfaces this question did not arise). We will adopt free boundary conditions for  $A_i$ , but require  $\varepsilon = 0$  on the boundary of  $S$ . (A rationale for this choice is that the Chern-Simons action is not invariant under gauge transformations that do not vanish on the boundary.) With this condition,  $\varepsilon$  generates in (5.2) the group  $\hat{G}_1$  of gauge transformations which are the identity on the boundary of the disc. An element of this group is an arbitrary continuous map  $V: D \rightarrow G$  whose restriction to  $S$  is the

identity. Now we impose the constraints. The first step is to require that  $\varepsilon^{ij} F_{ij}^a = 0$ . Since the disc is simply connected, this implies that  $A_i = -\partial_i U \cdot U^{-1}$  for some map  $U: D \rightarrow G$ .  $U$  is uniquely defined up to

$$U \rightarrow U \cdot W, \quad (5.3)$$

for a constant element  $W \in G$ . Then we must identify two  $U$ 's that differ by an element of the restricted gauge group  $\hat{G}_1$ . This means that we must impose the equivalence relation  $U \simeq VU$  for any  $V$  such that  $V = 1$  on the boundary  $S$ . The equivalence relation means that only the restriction of  $U$  to  $S$  is relevant. This restriction defines an element of the loop group  $LG$ , but because of the freedom (5.3), we should actually regard  $U$  as an element of  $LG/G$ . Geometrically, we have learned that the homogeneous space  $LG/G$  can be regarded as the symplectic quotient of the space of  $G$  connections on the disc  $D$ , by the group  $\hat{G}_1$  of symplectic diffeomorphisms. Now we wish to quantize the theory, which means doing quantum mechanics on  $LG/G$ , which inherits a natural symplectic structure from (5.1). Clearly, the group  $LG$  of gauge transformations on the boundary of  $\Sigma$  acts on  $LG/G$ , so the quantum Hilbert space will be at least a projective representation of the loop group. In fact, according to Segal [39], the quantization of  $LG/G$ , with this symplectic structure, gives rise to the basic irreducible highest weight representation of the loop group. This makes the connection between  $2+1$  dimensions and  $1+1$  dimensions far more direct, since the irreducible representations of the loop group are a basic ingredient in the  $1+1$  dimensional theory. It is obvious at this point that by considering more complicated Riemann surfaces with various boundary components, we can generate the whole  $1+1$  dimensional conformal field theory, essentially by studying the generally covariant  $2+1$  dimensional theory on various three manifolds with boundary.

*Acknowledgements.* This work originated with the realization that some results about conformal field theory described by G. Segal could be given a three dimensional interpretation by considering a gauge theory with Chern-Simons action. I am grateful to Segal for explaining his results, and to M. Atiyah for interesting me in and educating me about the Jones polynomial. V.F.R. Jones and L. Kauffman, and other participants at the IAMP Congress, raised many relevant questions. Finally, I must thank S. Deser and D.J. Gross for pointing out Polyakov's paper, G. Moore and N. Seiberg for explanations of their work, and the organizers of the IAMP Congress for their hospitality.

## References

1. Atiyah, M.F.: New invariants of three and four dimensional manifolds. In: The mathematical heritage of Hermann Weyl. Proc. Symp. Pure Math., vol. 48. Wells R. (ed.). Providence, RI: American Mathematical Society 1988, pp. 285–299
2. Donaldson, S.: An application of gauge theory to the topology of four manifolds. J. Diff. Geom. **18**, 269 (1983), Polynomial invariants for smooth four-manifolds. Oxford preprint
3. Floer, A.: An instanton invariant for three manifolds. Courant Institute preprint (1987). Morse theory for fixed points of symplectic diffeomorphisms. Bull. AMS **16**, 279 (1987)
4. Witten, E.: Topological quantum field theory. Commun. Math. Phys. **117**, 353 (1988)

5. Jones, V.F.R.: Index for subfactors. *Invent. Math.* **72**, 1 (1983). A polynomial invariant for links via von Neumann algebras. *Bull. AMS* **12**, 103 (1985), Hecke algebra representations of braid groups and link polynomials. *Ann. Math.* **126**, 335 (1987)
6. Freyd, P., Yetter, D., Hoste, J., Lickorish, W.B.R., Millett, K., Ocneanu, A.: A new polynomial invariant of knots and links. *Bull. AMS* **12**, 239 (1985)
7. Kauffman, L.: State models and the Jones polynomial. *Topology* **26**, 395 (1987). Statistical mechanics and the Jones polynomial, to appear in the Proceedings of the July, 1986, conference on Artin's braid group, Santa Cruz, California; An invariant of regular isotopy preprint
8. Turaev, V.G.: The Yang-Baxter equation and invariants of links. LOMI preprint E-3-87, *Inv. Math.* **92**, 527 (1988)
9. Przytycki, J.H., Traczyk, P.: Invariants of links of conway type. *Kobe J. Math.*, **4**, 115 (1988)
10. Birman, J.: On the Jones polynomial of closed 3-braids. *Invent. Math.* **81**, 287 (1985). Birman, J., Wenzl, H. Link polynomials and a new algebra, preprint
11. Tsuchiya, A., Kanie, Y.: In: Conformal field theory and solvable lattice models. *Adv. Stud. Pure math.* **16**, 297 (1988); *Lett. Math. Phys.* **13**, 303 (1987)
12. Verlinde, E.: Fusion rules and modular transformations in 2d conformal field theory. *Nucl. Phys.* **B300**, 360 (1988)
13. Moore, G., Seiberg, N.: Polynomial equations for rational conformal field theories. To appear in *Phys. Lett. B*, Naturality in conformal field theory. To appear in *Nucl. Phys. B*, Classical and quantum conformal field theory. IAS preprint HEP-88/35
14. Schroer, B.: *Nucl. Phys.* **295**, 4 (1988) K.-H. Rehren, Schroer, B.: Einstein causality and Artin braids. FU preprint (1987)
15. Fröhlich, J.: Statistics of fields, the Yang-Baxter equation, and the theory of links and knots. 1987 Cargèse lectures, to appear In *Nonperturbative quantum field theory*. New York: Plenum Press
16. Segal, G.: Conformal field theory. Oxford preprint; and lecture at the IAMP Congress, Swansea, July, 1988
17. Gromov, M.: Pseudo-holomorphic curves in symplectic manifolds. *Invent. Math.* **82**, 307 (1985)
18. Schwarz, A.: The partition function of degenerate quadratic functional and Ray-Singer invariants. *Lett. Math. Phys.* **2**, 247 (1978)
19. Schonfeld, J.: A mass term for three dimensional gauge fields. *Nucl. Phys.* **B185**, 157 (1981)
20. Jackiw, R., Templeton, S.: How superrenormalizable theories cure their infrared divergences. *Phys. Rev.* **D23**, 2291 (1981)
21. Deser, S., Jackiw, R., Templeton, S.: Three dimensional massive gauge theories. *Phys. Rev. Lett.* **48**, 975 (1983). Topologically massive gauge theory. *Ann. Phys. NY* **140**, 372 (1984)
22. Zuckerman, G.: Action principles and global geometry. In: The proceedings of the 1986 San Diego Summer Workshop, Yau, S.-T. (ed.)
23. Polyakov, A.M.: Fermi-bose transmutations induced by gauge fields. *Mod. Phys. Lett.* **A3**, 325 (1988)
24. Hagen, C.R.: *Ann. Phys.* **157**, 342 (1984)
25. Arovas, D., Schrieffer, R., Wilczek, F., Zee, A.: Statistical mechanics of anyons. *Nucl. Phys.* **B251[FS 13]**, 117 (1985)
26. Witten, E.: Non-Abelian bosonization in two dimensions. *Commun. Math. Phys.* **92**, 455 (1984)
27. Ray, D., Singer, I.: *Adv. Math.* **7**, 145 (1971), *Ann. Math.* **98** 154 (1973)
28. Deser, S., Jackiw, R., Templeton, S.: In [21]; Affleck, I., Harvey, J., Witten, E.: *Nucl. Phys.* **B206**, 413 (1982)  
Redlich, A.N.: Gauge invariance and parity conservation of three-dimensional fermions. *Phys. Rev. Lett.* **52**, 18 (1984)  
Alvarez-Gaumé, L., Witten, E.: Gravitational anomalies. *Nucl. Phys.* **B234**, 269 (1983)  
Alvarez-Gaumé, Della Pietra, S., Moore, G.: Anomalies and odd dimensions. *Ann. Phys. (NY)* **163**, 288 (1985)

- Atiyah, M.F.: A note on the eta invariant (unpublished)
- Singer, I.M.: Families of Dirac operators with applications to Physics. *Astérisque*, **1985**, p. 323
29. Atiyah, M.F., Patodi, V., Singer, I.: *Math. Proc. Camb. Phil. Soc.* **77**, 43 (1975), **78**, 405 (1975), **79**, 71 (1976)
30. Wilczek, F., Zee, A.: Linking numbers, spin, and statistics of solitons. *Phys. Rev. Lett.* **51**, 2250 (1983)
31. Friedan, D., Shenker, S.: *Nucl. Phys.* **B 281**, 509 (1987)
32. Belavin, A., Polyakov, A.M., Zamolodchikov, A.: *Nucl. Phys.* **B** (1984)
33. Atiyah, M.F., Bott, R.: The Yang-Mills equations over Riemann surfaces. *Phil. Trans. R. Soc. Lond. A* **308**, 523 (1982)
34. Quillen, D.: Determinants of Cauchy-Riemann operators over a Riemann surface. *Funct. Anal. Appl.* **19**, 31 (1986)
35. Drinfeld, V.: Quantum groups. In: *The Proceedings of the International Congress of Mathematicians, Berkeley 1986*, Vol. 1, pp. 798–820
36. Gepner, D., Witten, E.: String theory on group manifolds. *Nucl. Phys. B* **278**, 493 (1986)
37. Kac, V.G.: *Infinite dimensional Lie algebras*. Cambridge: Cambridge University Press (1985)
- Kac, V.G., Peterson, D.H.: *Adv. Math.* **53**, 125 (1984)
- Kac, V.G., Wakimoto, M.: *Adv. Math.* **70**, 156 (1988)
38. Witten, E.: 2+1 dimensional gravity as an exactly soluble system. IAS preprint HEP-88/32
39. Segal, G.: Unitary representation of some infinite dimensional groups. *Commun. Math. Phys.* **80**, 301 (1981)

Communicated by A. Jaffe

Received September 12, 1988; in revised form September 27, 1988

# 10. Chiral Anomalies In Field Theories

H. Banerjee \*

S. N. Bose National Centre for Basic Sciences,  
Salt Lake, Calcutta - 700 091, India

## Abstract

The role of the contribution from the fermion mass term in the axial vector Ward identity in generating the U(1) axial anomaly, both local and global, is elucidated. Gauge invariance requires the fermion to decouple from the gauge field if it is very heavy. This identifies the Adler-Bell-Jackiw (ABJ) anomaly with the asymptotic limit of the sign reversed mass term. In an instanton background, the chiral limit ( $m = 0$ ) of the mass term does not vanish but consists of contributions from fermion zero modes. Space time integral of these zero mode contributions exactly cancels, thanks to the Atiyah-Singer index theorem, the integral of the ABJ anomaly and suggests that the Jacobian for global U(1) chiral transformation is trivial even in an instanton background. This can be realised in the representation of the fermion partition function in a Weyl basis. The resolution of the strong CP problem is thus achieved in an axionless physical world.

In chiral gauge theories the fermion partition function admits of a gauge invariant representation but only at the cost of locality. Implementation of fermion averaging of the gauge current with the invariant partition function yields the current whose covariant derivative is the covariant anomaly. With the covariant current as input one can derive an integrable current whose covariant derivative is the minimal consistent anomaly obeying the Wess-Zumino consistency condition. The distinction between the two currents disappears if either the covariant or the consistent anomaly vanishes. This is realised only if the fermion belongs to an anomaly-free representation of the gauge group.

## 1 Introduction

In classical field theories there is a correspondence between a global symmetry of the action and a conserved Noether current. Presence of short distance singularities which need to be regularised for mathematical consistency complicates matters in quantum field theory (QFT). It may so happen that a regularisation scheme with mandatory attributes, like gauge invariance in a gauge theory of fermions, and at the same time consistent with the global symmetry cannot be formulated or simply does not exist. Traces of violation of the global symmetry in the form of non-conservation of the Noether current may survive as the regulator is removed at the end of calculation. This is the genesis of anomalies and anomalous Ward identities in QFT.

The topic of anomaly, in particular, axial anomaly came on the centrestage of particle physics research through the studies of neutral pion decay into two photons. The decay rate  $1.2 \times 10^{16}$  per sec. was explained satisfactorily by Steinberger<sup>1</sup> in 1949 in terms of triangle diagrams (Fig.1) with proton circulating in the fermion loop. The linear divergence of the amplitude was regulated by the Pauli-Villars method. Problem arose sixteen years later<sup>2</sup>, when decay rates obtained within the framework of current algebra and partial conservation of axial vector current (PCAC) were invariably smaller than the data by three orders of magnitude.

A popular working hypothesis, PCAC derives its dynamical basis in gauge theories of fermion like quantum chromodynamics (QCD) from the 'naive' operator relation (or, equivalently, naive Ward identity)

$$\partial_\mu (\bar{q} \gamma_5 \gamma_\mu \tau_3 q) = 2m (\bar{q} \gamma_5 \tau_3 q) \quad \dots (1.1)$$

---

\*Email: banerjee@boson.bose.res.in

which follows from field equations, with  $q$  the quark doublet ( $u, d$ ) and  $\tau_3$  the isospin. One recognises in the left hand side the Noether current corresponding to the chiral symmetry  $u \rightarrow e^{i\alpha\gamma_5}u, d \rightarrow e^{-i\alpha\gamma_5}d$ , which should be conserved at the classical level in the chiral limit  $m = 0$  of QCD. PCAC is just the statement that the mass term on the right hand side of (1.1) can be replaced by the neutral pion field

$$\partial_\mu (\bar{q}\gamma_5\gamma_\mu\tau_3q) = F_\pi m_\pi^2 \pi^0 \quad \dots (1.2)$$

where  $F_\pi$  is the pion decay constant,  $m_\pi$  the pion mass, and  $\pi^0$  the pion field. This is an unexceptionable step since the mass term has the right quantum numbers of a neutral pion and, therefore, can be regarded as the definition of the pion field in terms of quark constituents.

Problem with PCAC in  $\pi^0 \rightarrow 2\gamma$  stemmed from the Sutherland-Veltman<sup>3</sup> theorem which states that substitution of the divergence of isospin axial current for the neutral pion field in the matrix element yields a null result for the decay rate. Coupled with the positive result of Steinberger<sup>1</sup>, the unambiguous conclusion that emerges from the theorem is that the inadequacy of the PCAC relation stems really from the naive relation (1.1) which is flawed if quarks participate in electromagnetic interactions. The missing element was diagnosed as an anomaly, the Adler-Bell-Jackiw (ABJ) anomaly<sup>4</sup> in the Noether current for chiral symmetry

$$\partial_\mu (\bar{q}\gamma_5\gamma_\mu\tau_3q) = 2m (\bar{q}\gamma_5\tau_3q) - \left(\frac{N_c}{3}\right) \frac{e^2}{16\pi^2} \epsilon_{\mu\nu\alpha\beta} F_{\mu\nu} F_{\alpha\beta} \quad \dots (1.3)$$

where  $N_c$  is the colour degree of freedom of quarks and  $F_{\mu\nu}$  the electromagnetic field tensor. The ABJ anomaly, therefore, modifies the 'naive' PCAC relation (1.2) to

$$F_\pi m_\pi^2 \pi^0 = \partial_\mu (\bar{q}\gamma_5\gamma_\mu\tau_3q) + \left(\frac{N_c}{3}\right) \frac{e^2}{16\pi^2} \epsilon_{\mu\nu\alpha\beta} F_{\mu\nu} F_{\alpha\beta}$$

The decay rate now calculated by substituting the anomaly term for the pion field in the matrix element for  $\pi^0 \rightarrow 2\gamma$  is given by

$$\Gamma(\pi^0 \rightarrow 2\gamma) = \left(\frac{N_c}{3}\right)^2 \times 1.11 \times 10^{16} \text{ sec}^{-1} \quad \dots (1.4)$$

Depending on how one looks at the result (1.4), it may be regarded as either a signal success of the diagnosis of the problem in  $\pi^0 \rightarrow 2\gamma$  as due to anomaly, or in the light of later developments, a prediction of the number of colour degrees of freedom  $N_c = 3$  in QCD.

Success in  $\pi^0 \rightarrow 2\gamma$  problem brought into limelight the scenario of breaking symmetries at the classical level through anomalies in quantum field theories. Gauge theories become inconsistent if gauge symmetry is violated through anomaly. Cancellation of anomalies, therefore, constitutes an important constraint in building models for physical gauge theories with chiral coupling to fermions. Global chiral anomaly seems to play a key role in discussing physical effects associated with topologically nontrivial gauge field configurations.

## 2 Axial Anomaly and Fermion Decoupling

In a gauge theory of fermion there is a contradiction at the quantum level between chiral invariance and gauge symmetry. The ABJ anomaly, or in the more general context of non-Abelian gauge theories of fermion, the anomaly in the U(1) axial vector current, arises because gauge invariance is to be preserved for consistency of the theory. The contradiction is transparent in the condition for decoupling<sup>5</sup> of the fermion from the background gauge field when it is very heavy. For the divergence of the U(1) axial vector current the decoupling condition assumes the form of an anomalous Ward identity

$$\langle \partial_\lambda (\bar{\psi}(x)\gamma_5\gamma_\lambda\psi(x)) \rangle = 2m \langle \bar{\psi}(x)\gamma_5\psi(x) \rangle - \lim_{m \rightarrow \infty} [2m \langle \bar{\psi}(x)\gamma_5\psi(x) \rangle] \quad \dots (2.1)$$

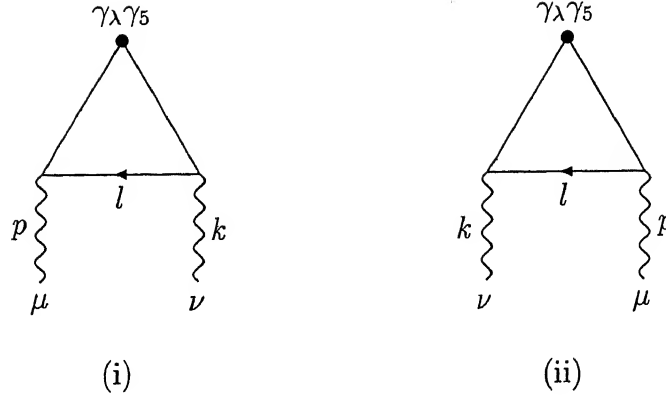


Figure 1: Triangle diagrams

where  $\langle \rangle$  denotes that the fermion degrees of freedom are integrated out. As we shall see below, (2.1) follows directly from gauge invariance and is known as Adler<sup>6</sup> condition in QED. Eq.(2.1) will still be compatible with chiral symmetry and a conserved U(1) axial vector current would emerge in the chiral limit  $m = 0$  if the second term on the right hand side were to vanish. But this is not to be. The asymptotics in field theory gives in the infinite mass limit the ABJ anomaly

$$\lim_{m \rightarrow \infty} [2m \langle \bar{\psi}(x) \gamma_5 \psi(x) \rangle] = \frac{g^2}{16\pi^2} \epsilon_{\mu\nu\lambda\rho} \text{tr} F_{\mu\nu} F_{\lambda\rho} \quad \dots (2.2)$$

where  $F_{\mu\nu} = F_{\mu\nu}^a t_a$  are the field tensors with  $t_a$  the generators of the gauge group.

To motivate the decoupling condition (2.1) we consider in QED the amplitude for creating two photons with momenta and polarisation  $(p, \mu)$  and  $(k, \nu)$  by the axial vector current  $J_{\lambda 5}(x) = \bar{\psi}(x) \gamma_5 \gamma_\lambda \psi(x)$

$$\langle p, \mu; k, \nu | J_{\lambda 5}(0) | 0 \rangle = \epsilon_\mu(p) \epsilon_\nu(k) M_{\lambda\mu\nu}(p, k, m)$$

The key to the analysis is the Rosenberg<sup>7</sup> tensor decomposition (see Fig.1)

$$M_{\lambda\mu\nu} = \epsilon_{\lambda\mu\nu\alpha} k_\alpha A(p, k, m) + \epsilon_{\lambda\nu\alpha\beta} p_\alpha k_\beta [p_\mu B(p, k, m) + k_\mu C(p, k, m)] + [(k, \nu) \leftrightarrow (p, \mu)] \quad \dots (2.3)$$

which follows from parity and Lorentz invariance.

Not all the form factors are independent. The form factor  $A$  which gives the divergence of the axial vector current

$$(p + k)_\lambda M_{\lambda\mu\nu} = \epsilon_{\mu\nu\alpha\beta} p_\alpha k_\beta [A(p, k, m) + A(k, p, m)] \quad \dots (2.4)$$

is determined through gauge invariance by the form factors  $B$  and  $C$

$$A(p, k, m) = p^2 B(p, k, m) + p \cdot k C(p, k, m) \quad \dots (2.5)$$

The form factors  $B$  and  $C$  are of dimensions  $[\text{mass}]^{-2}$ , and, therefore, in perturbation theory they are represented by highly convergent amplitudes which vanish as  $m^{-2}$  for large fermion mass

$$\lim_{m \rightarrow \infty} B(p, k, m) = \lim_{m \rightarrow \infty} C(p, k, m) = 0$$

Thus gauge invariance (2.5) guarantees that the divergence of the amplitude for the axial vector current given in (2.4) vanishes in the asymptotic  $m \rightarrow \infty$  limit

$$\lim_{m \rightarrow \infty} (p+k)_\lambda M_{\lambda\mu\nu} = 0 \quad \dots (2.6)$$

In perturbation theory the amplitude  $M_{\lambda\mu\nu}$  for the triangle diagram is linearly divergent. The leading divergence, however, drops out due to symmetric integration of loop momentum leaving a potential logarithmic divergence, which can appear only in the form factor  $A$  in (2.3). Gauge invariance (2.5) rules out even this residual logarithmic divergence.

The above observations suggest that the potential anomaly represented by the second term in (2.1) must be finite and independent of regularisation scheme. To verify this we start by calculating

$$\lim_{m \rightarrow \infty} [2m \langle \bar{\psi}(x) \gamma_5 \gamma(x) \rangle] = \lim_{m \rightarrow \infty} [2m \langle x | \text{Tr} \gamma_5 (i\mathcal{D} + m)^{-1} | x \rangle], \quad \dots (2.7)$$

where, to conform to our discussions in the subsequent sections, we work in Euclidean metric and write for the hermitian Dirac operator

$$\mathcal{D} = \gamma_\mu (i\partial_\mu - gA_\mu) \quad \dots (2.8)$$

with  $A_\mu \equiv A_\mu^a t_a$ , the gauge potential.

Our strategy is to develop the Green function appearing in (2.7) in a perturbative series

$$(i\mathcal{D} + m)^{-1} = (-i\mathcal{D} + m)G$$

with

$$\begin{aligned} G &= (\not{p}^2 + m^2)^{-1} \\ &= G_0 - gG_0 V G_0 + g^2 G_0 V G_0 V G_0 + \dots \end{aligned} \quad \dots (2.9)$$

The ‘free’ part is  $G_0 = (p^2 + m^2)^{-1}$  with  $p_\mu = i\partial_\mu$  the ‘momentum’ operator. The ‘potential’  $gV$  has two pieces

$$gV = gV_0 + \frac{1}{2} \sigma_{\mu\nu} F_{\mu\nu}$$

with  $\sigma_{\mu\nu} = \frac{1}{2i} (\gamma_\mu \gamma_\nu - \gamma_\nu \gamma_\mu)$ . The first piece  $gV_0$  is at most linear in  $p$  and independent of  $\gamma$ -matrices. The trace with  $\gamma_5$  in (2.7) starts to be nonvanishing only from terms of order  $g^2$  onwards in the perturbative expansion (2.9) and one obtains (2.2)

$$\lim_{m \rightarrow \infty} [2m \langle \bar{\psi}(x) \gamma_5 \psi(x) \rangle] = \lim_{m \rightarrow \infty} [2m^2 \langle x | \text{Tr} (\gamma_5 G) | x \rangle] = \frac{g^2}{16\pi^2} \epsilon_{\mu\nu\lambda\rho} \text{tr} F_{\mu\nu}(x) F_{\lambda\rho}(x)$$

Note that the final result is local. All nonlocalities as well as contributions from higher order terms in the perturbative series (2.9) drop out in the infinite mass limit  $m \rightarrow \infty$ .

In the decoupling condition (2.1) one can set the mass term on the right hand side to zero in the chiral limit  $m = 0$ . The anomalous Ward identity thus obtained

$$\langle \partial_\lambda (\bar{\psi} \gamma_5 \gamma_\lambda \psi) \rangle_{m=0} = -\frac{g^2}{16\pi^2} \epsilon_{\mu\nu\lambda\rho} \text{tr} F_{\mu\nu} F_{\lambda\rho} \quad \dots (2.10)$$

shows that the U(1) axial vector current, i.e. the Noether current corresponding to global chiral symmetry

$$\psi \rightarrow e^{i\alpha\gamma_5} \psi, \quad \bar{\psi} \rightarrow \bar{\psi} e^{i\alpha\gamma_5}, \quad \dots (2.11)$$

of the massless Dirac operator (2.8), is not conserved. The divergence of the current is just the ABJ anomaly which is responsible for the two photon decay of neutral pion discussed in the preceding section.

We note that in renormalisable theories in perturbative framework the decoupling condition (2.1), which is a special example of the decoupling theorem of Appelquist and Carazzone<sup>5</sup>, is correct to all orders of perturbation, just as the Adler–Bardeen<sup>8</sup> theorem assures us that the anomalous axial vector Ward identity (2.10) is not affected by radiative corrections in QED.



It should be remarked that setting the mass term to zero in (2.1), as was done in obtaining (2.10), may not always be legitimate in the chiral limit  $m = 0$  if the gauge field is treated nonperturbatively. The Euclidean Dirac operator (2.8) has zero modes if the background gauge field has a nontrivial topology. In this scenario the chiral limit of the mass term does not vanish and, as we shall see in the next section, consists precisely of the zero modes of the Dirac operator.

### 3 Path Integral Approach to Anomaly

In a seminal work Fujikawa<sup>9</sup> interpreted the ABJ anomaly within the path integral framework as arising from the nontrivial Jacobian of the fermion measure under chiral transformation. In Euclidean metric, considered by Fujikawa, the partition function which generates fermion Green functions in a background gauge field configuration may be written as

$$Z_f[A] \equiv \int d\mu \exp \left[ \int \bar{\psi}(i\mathcal{D} + m)\psi d^4x \right] \quad \dots (3.1)$$

where  $d\mu$  is the integration measure for fermion and  $\mathcal{D}$  is the Euclidean Dirac operator defined in (2.8). Fermion Green functions which are normalised expectation values of any product  $O$  of fermion fields are obtained from the functional integral (3.1)

$$\langle O \rangle = \frac{1}{Z_f[A]} \int d\mu O \exp \left[ \int \bar{\psi}(i\mathcal{D} + m)\psi d^4x \right] \quad \dots (3.2)$$

With  $\gamma$ -matrices chosen hermitian, the Dirac operator (2.8) is also hermitian with real eigenvalues  $\lambda_n$  and orthonormal eigenfunctions  $\phi_n(x)$

$$\mathcal{D}\phi_n(x) = \lambda_n \phi_n(x), \quad \int \phi_m^\dagger(x) \phi_n(x) d^4x = \delta_{mn} \quad \dots (3.3)$$

Each nonzero eigenvalue  $\lambda_n$  has its chirally conjugate partner  $-\lambda_n$  with eigenfunctions  $\phi_{-n}$

$$\mathcal{D}\phi_{-n} = -\lambda_n \phi_{-n}, \quad \phi_{-n} = \gamma_5 \phi_n \quad \dots (3.4)$$

In perturbative field theories one is interested in gauge field configurations with only trivial topology. For such configurations the kernel space of the Euclidean Dirac operator (2.8) is of dimension zero. This means that the set  $\{\phi_n(x)\}$  with nonzero eigenvalues constitute a complete basis in function space. The Dirac field  $\psi(x)$  can be expanded in this basis as

$$\psi(x) = \sum (a_n + a_{-n}\gamma_5) \phi_n(x) \quad \dots (3.5)$$

where  $a_{\pm n}$  are complex-valued Grassmann generators. The four degrees of freedom corresponding to each mode of the Dirac field is accounted for if we split  $a_{\pm n}$  as

$$a_{\pm n} = \alpha_{\pm n} + i\beta_{\pm n}$$

with  $\alpha, \beta$  real valued.

There are ambiguities<sup>10,11,12</sup> on the issue whether in Euclidean metric  $\bar{\psi}$  should be treated as independent of  $\psi$ . For the present, we follow the popular ansatz<sup>10</sup> and expand  $\bar{\psi}(x)$  with an independent set of Grassmann generators  $\{\bar{b}_{\pm n}\}$ ,

$$\bar{\psi}(x) = \sum \phi_n^\dagger(x) (\bar{b}_n + \bar{b}_{-n}\gamma_5) \quad \dots (3.6)$$

The representations for  $\psi(x)$  and  $\bar{\psi}(x)$  together with orthonormality of the eigenfunctions (3.3) yield for the fermion action

$$\begin{aligned} S_f(m) &= \int \bar{\psi}(x)(i\mathcal{D} + m)\psi(x) d^4x \\ &= \sum [(i\lambda_n + m) \bar{b}_n a_n + (-i\lambda_n + m) \bar{b}_{-n} a_{-n}] \end{aligned} \quad \dots (3.7)$$

The integration measure for the fermion fields in the basis  $\{\phi_n(x)\}$  is

$$d\mu \equiv \Pi_n d\bar{b}_n da_n d\bar{b}_{-n} da_{-n} \quad \dots (3.8)$$

The standard rules of integration of Grassmann generators now yield for the partition function (3.1) the desired result

$$\begin{aligned} Z_f[A] &= \Pi_n (\lambda_n^2 + m^2) \\ &= \det(i\mathcal{D} + m) \end{aligned} \quad \dots (3.9)$$

This confirms the correctness of the choice of the measure (3.8).

Ward identities, whether normal or anomalous, are obtained in path integral framework from the requirement that the partition function is invariant under infinitesimal symmetry transformations of the variables of integration. To derive the Ward identity corresponding to chiral symmetry one implements a 'local' chiral transformation of the variables of integration  $\psi(x), \bar{\psi}(x)$  in the partition function (3.1)

$$\begin{aligned} \psi(x) &\rightarrow \psi'(x) = (1 + i\alpha(x)\gamma_5) \psi(x) \\ \bar{\psi}(x) &\rightarrow \bar{\psi}'(x) = \bar{\psi}(x) (1 + i\alpha(x)\gamma_5) \end{aligned} \quad \dots (3.10)$$

The fermion measure (3.8) changes and the new measure corresponding to the transformed variables of integration is given by  $d\mu'$

$$\begin{aligned} d\mu' &= \Pi_n d\bar{b}'_n da'_n d\bar{b}'_{-n} da'_{-n} \\ &= d\mu J[\alpha], \end{aligned} \quad \dots (3.11)$$

where  $a'_n(\bar{b}'_n)$  are the new set of Grassmann generators in the expansion of  $\psi'(x) (\bar{\psi}'(x))$  in the basis  $\{\phi_n(x)\}$ . The Jacobian  $J[\alpha]$  can be calculated following standard procedure

$$J[\alpha] = \exp \left[ -2i \int d^4x \alpha(x) \sum_n (\phi_n^+(x) \gamma_5 \phi_n(x) + \phi_{-n}^+(x) \gamma_5 \phi_{-n}(x)) \right] \quad \dots (3.12)$$

The fermion action also changes and the new action is given by

$$S_f(m) \rightarrow S'_f(m) = \int d^4x [\bar{\psi}(i\mathcal{D} + m)\psi + i\alpha(x) (-\partial_\mu (\bar{\psi}\gamma_5\gamma_\mu\psi) + 2m\bar{\psi}\gamma_5\psi)] \quad \dots (3.13)$$

Invariance of the partition function (3.1) under the infinitesimal local chiral transformation (3.10) now gives the anomalous axial Ward identity

$$\langle \partial_\mu (\bar{\psi}\gamma_5\gamma_\mu\psi) \rangle = 2m \langle \bar{\psi}\gamma_5\psi \rangle - 2 \sum_n (\phi_n^+ \gamma_5 \phi_n + \phi_{-n}^+ \gamma_5 \phi_{-n}) \quad \dots (3.14)$$

It is easy to recognise (3.14) as the decoupling condition (2.1) of the preceding section. Indeed

$$\begin{aligned} \lim_{m \rightarrow \infty} [2m \langle \bar{\psi}\gamma_5\psi \rangle] &= \lim_{m \rightarrow \infty} \left[ 2m \sum_n \left( \frac{\phi_n^+ \gamma_5 \phi_n}{m + i\lambda_n} + \frac{\phi_{-n}^+ \gamma_5 \phi_{-n}}{m - i\lambda_n} \right) \right] \\ &= 2 \sum_n (\phi_n^+ \gamma_5 \phi_n + \phi_{-n}^+ \gamma_5 \phi_{-n}) \end{aligned} \quad \dots (3.15)$$

where the infinite sum on the right hand side of (3.14) is to be cut off gauge invariantly,  $|\lambda_n| \ll M$  for  $M$  large. Fujikawa used the gauge invariant cut off  $\exp(-\mathcal{D}^2/M^2)$  with large  $M$  to evaluate the infinite sum

$$\begin{aligned} 2 \sum_n (\phi_n^+(x) \gamma_5 \phi_n(x) + \phi_{-n}^+(x) \gamma_5 \phi_{-n}(x)) &= \lim_{M \rightarrow \infty} \langle x | \text{Tr} \left( \gamma_5 e^{-\mathcal{D}^2/M^2} \right) | x \rangle \\ &= \frac{g^2}{16\pi^2} \epsilon_{\mu\nu\lambda\rho} \text{tr} F_{\mu\nu}(x) F_{\lambda\rho}(x) \end{aligned} \quad \dots (3.16)$$

It is clear that the left hand side of (3.16) should be augmented by zero modes if the Dirac operator admits of them. Zero modes always appear with definite chiralities,  $\epsilon_i = \pm 1$

$$\mathcal{D} \phi_{0i} = 0, \quad \gamma_5 \phi_{0i} = \epsilon_i \phi_{0i} \quad \dots (3.17)$$

This is because, in its kernel space the Dirac operator commutes with  $\gamma_5$ . In the presence of zero modes the left hand side of (3.16) needs to be augmented by their contributions, i.e.,

$$\frac{g^2}{16\pi^2} \epsilon_{\mu\nu\lambda\rho} \text{tr} F_{\mu\nu}(x) F_{\lambda\rho}(x) = 2 \left[ \sum_i \epsilon_i \phi_{0i}^+ \phi_{0i} + \sum_n (\phi_n^+ \gamma_5 \phi_n + \phi_{-n}^+ \gamma_5 \phi_{-n}) \right] \quad \dots (3.18)$$

Space time integral of (3.18) gives the Atiyah–Singer<sup>13</sup> index theorem

$$\nu \equiv \frac{g^2}{32\pi^2} \epsilon_{\mu\nu\lambda\rho} \int d^4x \text{tr} F_{\mu\nu}(x) F_{\lambda\rho}(x) = n_+ - n_- \quad \dots (3.19)$$

where  $\nu$  is the winding number (Pontryagin index) of the gauge field and  $n_+$  ( $n_-$ ) is the number of positive (negative) chirality zero modes. Eigenmodes corresponding to nonzero eigenvalues do not contribute to the space time integral (3.19) because  $\phi_n$  is orthogonal to  $\gamma_5 \phi_n$ . Note that nontrivial winding number,  $\nu \neq 0$ , is realised through instanton-like configuration of the gauge field.

The presence of zero modes has profound impact on the chiral limit of the fermion mass term on the right hand side of the anomalous axial Ward identity

$$\langle \partial_\mu (\bar{\psi} \gamma_5 \psi) \rangle = 2m \langle \bar{\psi} \gamma_5 \psi \rangle - \frac{g^2}{16\pi^2} \epsilon_{\mu\nu\lambda\rho} \text{tr} F_{\mu\nu} F_{\lambda\rho} \quad \dots (3.20)$$

The zero modes can be isolated from the mass term

$$2m \langle \bar{\psi} \gamma_5 \psi \rangle = 2m \langle \bar{\psi} \gamma_5 \psi \rangle' + 2 \sum \epsilon_i \phi_{0i}^+ \phi_{0i} \quad \dots (3.21)$$

where the first term on the right hand side is bereft of the zero modes and vanishes in the chiral limit

$$2m \langle \bar{\psi} \gamma_5 \psi \rangle' = 4 \sum_{\lambda_n > 0} \frac{m^2}{m^2 + \lambda_n^2} \phi_n^+ \gamma_5 \phi_n \quad \dots (3.22)$$

The mass term, therefore, has now a nontrivial chiral limit consisting precisely of the zero modes, and the chiral limit of the axial Ward identity is not exactly what was obtained in the perturbative framework of the preceding section

$$\langle \partial_\mu (\bar{\psi} \gamma_5 \gamma_\mu \psi) \rangle_{m=0} = 2 \sum \epsilon_i \phi_{0i}^+ \phi_{0i} - \frac{g^2}{16\pi^2} \epsilon_{\mu\nu\lambda\rho} \text{tr} F_{\mu\nu} F_{\lambda\rho} \quad \dots (3.23)$$

The zero modes in the extra piece appearing on the right hand side arise from instanton-like configuration of the gauge field and, therefore, could not have been accessed in a perturbative framework.

It is of great interest to note that the zero mode terms which appear explicitly on the right hand side of (3.23) are exactly cancelled by similar terms contained now (see (3.18)) in the ABJ anomaly. Thus, irrespective of whether or not the gauge field configuration gives rise to zero modes, the chiral limit of the local axial anomaly comprises of only nonzero eigenmodes of the Dirac operator

$$\langle \partial_\mu (\bar{\psi} \gamma_5 \gamma_\mu \psi) \rangle_{m=0} = -2 \sum_{|\lambda_n| \neq 0} (\phi_n^+ \gamma_5 \phi_n + \gamma_{-n}^+ \gamma_5 \phi_{-n}) \quad \dots (3.24)$$

We, therefore, conclude that the space-time integral of the chiral limit of the divergence of the axial vector current always vanishes. This follows from the orthogonality of  $\phi_n$  and  $\gamma_5 \phi_n$  if one uses (3.24) or from the Atiyah–Singer index theorem if instead one uses (3.23)

$$\int \langle \partial_\mu (\bar{\psi} \gamma_5 \gamma_\mu \psi) \rangle_{m=0} d^4x = 0 \quad \dots (3.25)$$

This, as we shall see later, has a profound impact on issues of physics related to global chiral anomaly.

## 4 Chiral Gauge Theories and the Covariant and Consistent Anomalies

The ABJ anomaly in the U(1) axial vector Ward identity constitutes an unambiguous evidence of a fundamental incompatibility of chiral invariance and gauge symmetry in regularisation scheme in perturbative framework of quantum field theory. In a vector-like gauge theory, such as QCD, chiral invariance is an expendable attribute and the ABJ anomaly results from strict adherence to gauge symmetry. In chiral gauge theories where gauge fields are coupled chirally to fermions in the Dirac operator

$$\begin{aligned}\mathcal{D} &\equiv \gamma_\mu (i\partial_\mu + t_a A_\mu^a \frac{1}{2}(1 - \gamma_5)) \\ &= (i\cancel{\partial} + A \frac{1}{2}(1 - \gamma_5)),\end{aligned}\quad \dots (4.1)$$

loss of chiral invariance jeopardises gauge symmetry and hence the consistency of the theory. The fermion action

$$S_F = \int \bar{\psi} \mathcal{D} \psi d^4x \quad \dots (4.2)$$

is invariant under the local chiral gauge transformations

$$\begin{aligned}\psi(x) &\rightarrow e^{i\alpha(x)\frac{(1-\gamma_5)}{2}}\psi(x), \quad \bar{\psi}(x) \rightarrow \bar{\psi}(x)e^{-i\alpha(x)\frac{(1+\gamma_5)}{2}} \\ A_\mu(x) &\rightarrow e^{i\alpha(x)} \left\{ A_\mu(x) + \frac{1}{i}\partial_\mu \right\} e^{-i\alpha(x)}\end{aligned}\quad \dots (4.3)$$

with  $\alpha(x) = t_a \alpha_a(x)$  the gauge function. Dimensional regularisation, popular in perturbative gauge theories, has serious problem with  $\gamma_5 = \frac{1}{4!}\epsilon_{\mu\nu\lambda\rho}\gamma_\mu\gamma_\nu\gamma_\lambda\gamma_\rho$ . The totally antisymmetric tensor of rank four  $\epsilon_{\mu\nu\lambda\rho}$  does not admit of suitable generalisation to arbitrary space-time dimensions. Thus, one is yet to find a consistent and systematic scheme for regulating divergences in chiral gauge theories in weak coupling perturbation in the continuum.

On lattice, the finite spacing  $a$  between lattice sites provides a built-in regularisation of all short distance singularities in field theories. Here too, the prospects for a consistent formulation of chiral gauge theory are not really bright. The major problem on lattice is the species doublers of fermion and their removal. The doublers appear as unwanted zeros of the Fourier transform of the ‘free’ Dirac operator on lattice, over and above the zero at the origin of momentum space which correspond to the physical fermion. In the ‘naive’ Dirac operator  $(\gamma_\mu \sin(p_\mu a)/a)$  the doublers are located at the edges of the Brillouin zone  $-(\pi/a) \leq p_\mu \leq (\pi/a)$ . The doublers are not specific for the naive Dirac operator. According to the celebrated theorem of Nielsen and Ninomiya<sup>14</sup> these are generic and can be avoided only at a price, by breaking explicitly locality and/or chiral symmetry in the Dirac operator. The most popular model for lattice fermion, the Wilson model<sup>15</sup>, removes the doublers by giving them masses of the order of the lattice cut-off  $O(1/a)$

$$D_W(p) = \gamma_\mu \sin(p_\mu a)/a + i\tau(1 - \cos(p_\mu a))/a \quad \dots (4.3)$$

Gauge invariance is implemented simply through link variables as in all lattice models. But the explicit breaking of chiral symmetry for nonzero ‘ $\tau$ ’ makes the model patently inappropriate for chiral gauge theories. Current spurt in interest in the subject stems mainly from the realisation that for lattice Dirac operators  $D$  obeying the Ginsparg–Wilson<sup>16</sup> relation

$$\gamma_5 D + D \gamma_5 = a D \gamma_5 D, \quad \dots (4.4)$$

chiral symmetry is restored and species doublers are removed in the continuum limit<sup>17</sup>. The issue of nonlocality implied in the Ginsparg–Wilson relation, particularly in the context of chiral gauge theories, is yet to be resolved<sup>18</sup>.

**Covariant Anomaly**<sup>19</sup>: Apart from the absence of a consistent and systematic regularisation scheme, chiral gauge theories are, in general, afflicted with anomalies in the gauge current. The Dirac operator (4.1) in chiral gauge theory is non-hermitian. A fallout of this is that Fujikawa’s<sup>9</sup> recipe for constructing a gauge invariant partition function, which assumes a hermitian Dirac

operator, needs to be modified. The Dirac operator  $\mathcal{D}$  in (4.1) maps  $\psi$  into the space of spinors in the domain of  $\mathcal{D}^+$ . The eigenvalue equations (3.3) are, therefore replaced by

$$\mathcal{D}\phi_n = \lambda_n\chi_n, \quad \mathcal{D}^+\chi_n = \lambda_n\phi_n, \quad \dots (4.5)$$

where  $\lambda_n^2$  are real, nonnegative and constitute the eigenvalue spectrum of  $\mathcal{D}\mathcal{D}^+$  and  $\mathcal{D}^+\mathcal{D}$ . The sets of eigen functions  $\{\phi_n\}$  and  $\{\chi_n\}$  of  $\mathcal{D}^+\mathcal{D}$  and  $\mathcal{D}\mathcal{D}^+$  respectively constitute an orthonormal basis for expanding  $\psi$  and  $\bar{\psi}$

$$\psi = \sum_n a_n \phi_n, \quad \bar{\psi} = \sum_n \bar{b}_n \chi_n^+ \quad \dots (4.6)$$

in terms of the Grassmann generators  $a_n, \bar{b}_n$ . The fermion measure defined as

$$d\mu[A] = \Pi_n d\bar{b}_n da_n \quad \dots (4.7)$$

is a gauge invariant functional of  $A_\mu$  and yields the partition function<sup>19</sup>

$$\begin{aligned} Z_{inv}[A] &\equiv \int d\mu[A] \exp[\int \bar{\psi} \mathcal{D} \psi d^4x] \\ &= (\det \mathcal{D}^+ \mathcal{D})^{1/2} = (\det \mathcal{D} \mathcal{D}^+)^{1/2} \end{aligned} \quad \dots (4.8).$$

Both  $\mathcal{D}^+\mathcal{D}$  and  $\mathcal{D}\mathcal{D}^+$  change by a similarity transformation under gauge transformation. The representation (4.8) is thus formally gauge invariant.

The chiral gauge current

$$\left[ \bar{\psi} t_a \gamma_\mu \frac{1}{2} (1 - \gamma_5) \psi \right] \quad \dots (4.9)$$

transforms covariantly under gauge transformation (4.3). Fermion averaging of the current with the gauge invariant measure (4.7) yields

$$\begin{aligned} J_\mu^a(x) &\equiv \frac{\int d\mu[A] (\bar{\psi} t_a \gamma_\mu \frac{1}{2} (1 - \gamma_5) \psi) \exp[\int \bar{\psi} \mathcal{D} \psi d^4x]}{\int d\mu[A] \exp[\int \bar{\psi} \mathcal{D} \psi d^4x]} \\ &= \sum_n \frac{1}{\lambda_n} \chi_n^+ t_a \gamma_\mu \frac{1}{2} (1 - \gamma_5) \phi_n. \end{aligned} \quad \dots (4.10)$$

Gauge invariant regularisation can be implemented by suppressing large eigenvalues and the current thus obtained transforms covariantly and is called the covariant current.

Formal application of field equations suggest that the gauge current should be covariantly conserved. This, however, may not be true for the fermion averaged current  $J_\mu^a(x)$  if it is anomalous,

$$\begin{aligned} G^a(x) &\equiv \partial_\mu J_\mu^a(x) - f^{abc} A_\mu^b(x) J_\mu^c(x) \\ &= \sum_n \{ \chi_n^+ t_a \frac{1}{2} (1 + \gamma_5) \chi_n - \phi_n^+ t_a \frac{1}{2} (1 - \gamma_5) \phi_n \} \end{aligned} \quad \dots (4.11)$$

Following Fujikawa's<sup>9</sup> recipe for gauge invariant regularisation one obtains the covariant anomaly

$$\begin{aligned} G^a(x) &= \lim_{M \rightarrow \infty} \int \frac{d^4k}{(2\pi)^4} \text{Tr} t_a \left[ \frac{1}{2} (1 + \gamma_5) e^{ik \cdot x} e^{-\frac{p \cdot p^+}{M^2}} e^{-ik \cdot x} \frac{1}{2} (1 + \gamma_5) e^{ik \cdot x} e^{-\frac{p \cdot p^+}{M^2}} e^{-ik \cdot x} \right] \\ &= -\frac{1}{32\pi^2} \epsilon_{\mu\nu\lambda\rho} \text{tr} [t_a F_{\mu\nu} F_{\lambda\rho}] \end{aligned} \quad \dots (4.12)$$

where  $F_{\mu\nu} = t_a F_{\mu\nu}^a$  are the field tensors.

**Consistent Anomaly**<sup>19,20</sup>: In perturbative treatment of chiral gauge theories the fermion measure in the partition function is independent of the gauge field. A fallout of this is that, unlike  $Z_{inv}[A]$  in (4.8), the perturbative partition function

$$\begin{aligned} Z_{pert}[A] &\equiv e^{W[A]} \\ &= \int d\mu \exp[\int \bar{\psi} \mathcal{D} \psi d^4x], \end{aligned} \quad \dots (4.13)$$

and hence the effective action  $W[A]$  need not be gauge invariant. The gauge current with fermion averaging implemented through this perturbative partition function

$$\begin{aligned} J_{W\mu}^a(x) &\equiv \frac{\delta}{\delta A_\mu^a(x)} W[A] \\ &= \langle \bar{\psi} t_a \gamma_\mu \frac{1}{2} (1 - \gamma_5) \psi \rangle_W \end{aligned} \quad \dots (4.14)$$

will, in general, not transform covariantly. However, it must obey the integrability condition

$$\frac{\delta J_{W\mu}^a(x)}{\delta A_\nu^b(x')} - \frac{\delta J_{W\nu}^b(x')}{\delta A_\mu^a(x)} = 0, \quad \dots (4.15)$$

since it is defined in (4.14) through the functional derivative of the effective action  $W[A]$ . The current  $J_{W\mu}^a(x)$  is called the consistent current and its covariant derivative

$$G^a(x) \equiv \partial_\mu J_{W\mu}^a(x) - f^{abc} A_\mu^b J_{W\mu}^c(x) \quad \dots (4.16)$$

is the consistent anomaly.

Gauge transformation properties of an arbitrary functional of gauge fields are best discussed with the help of the generators

$$L^a(x) = \partial_\mu \frac{\delta}{\delta A_\mu^a(x)} - f^{abc} A_\mu^b(x) \frac{\delta}{\delta A_\mu^c(x)} \quad \dots (4.17)$$

Thus the consistent anomaly  $G_W^a(x)$ , representing as it does the gauge variation of the effective action  $W[A]$ , is given by

$$G_W^a(x) = L^a(x) W[A] \quad \dots (4.18)$$

The algebra of the generators

$$[L^a(x), L^b(x')] = f^{abc} \delta^4(x - x') L^c(x) \quad \dots (4.19)$$

shows that the consistent anomaly must obey the Wess-Zumino<sup>21</sup> consistency condition

$$L^a(x) G_W^b(x') - L^b(x') G_W^a(x) = f^{abc} \delta^4(x - x') G_W^c(x) \quad \dots (4.20)$$

On the other hand, the anomaly  $G_W^a(x)$  is a measure of the non-covariance of the consistent current  $J_{W\mu}^a(x)$

$$L^b(x') J_\mu^a(x) = -f^{abc} \delta^4(x - x') + \frac{\delta G_W^b(x')}{\delta A_\mu^a(x)} \quad \dots (4.21)$$

As for the covariant anomaly (4.12), one finds, as expected, an incompatibility with the Wess-Zumino consistency condition

$$L^a(x) G^b(x') - L^b(x') G^a(x) = 2f^{abc} \delta^4(x - x') G^c(x), \quad \dots (4.22)$$

where the factor 2 on the right hand side spoils consistency. Thus, the anomaly itself is a measure of the 'inconsistency'. The origin of the 'inconsistency' may be traced to the fermion measure  $d\mu[A]$  given by (4.7) for averaging of the gauge current in the definition (4.10) of the covariant current  $J_\mu^a(x)$ . A nontrivial covariant anomaly  $G^a(x)$  corresponds to a nontrivial dependence of the measure  $d\mu[A]$  on the gauge field. This is suggested also from the observation that the definition

$$\mathcal{J}_\mu^a(x) \equiv \frac{\delta}{\delta A_\mu^a(x)} \ln Z_{inv}[A] \quad \dots (4.23)$$

where  $Z_{inv}[A]$  is the gauge invariant partition function (4.8), has all the attributes, it is covariant, consistent and anomaly free. The price that one pays for this 'perfect' current is a high degree of nonlinearity.

It can be shown<sup>19</sup> that the consistent current coincides with the covariant current if the functional curl of the latter vanishes

$$J_{W\mu}^a(x) = J_\mu^a(x) + \int_0^1 dg \int d^4x' A_\nu^b(x') \left\{ \frac{\delta J_\nu^{bg}(x')}{\delta A_\mu^a(x)} - \frac{\delta J_\mu^{ag}(x)}{\delta A_\nu^b(x')} \right\} \quad \dots (4.24)$$

where  $J_\mu^{ag}(x)$  is the covariant current corresponding to the Dirac operator  $\mathcal{D}^g = (i\partial + gA_{\frac{1}{2}}(1 - \gamma_5))$  with coupling constant  $g$ . One can obtain from (4.24) an explicit representation of the consistent anomaly using the expression (4.12) for the covariant anomaly<sup>19</sup>

$$G_W^a(x) = f_0^1 dg G^{ag} + \frac{1}{16\pi^2} \epsilon_{\mu\nu\lambda\rho} \int_0^1 dgg(1-g) \text{tr} \left( [t_a, A_\mu] \left( F_{\lambda\rho}^g A_\nu + A_\nu F_{\lambda\rho}^g \right) \right) \quad \dots (4.25)$$

The above analysis shows that the distinction between covariant and consistent currents disappears if and only if the anomaly in either current vanishes. The fundamental requirement that the chiral gauge theory is free of either anomaly imposes the unique constraint on the group generators of the chiral fermions

$$\text{tr} (t_a \{t_b, t_c\}) = 0 \quad \dots (4.26)$$

which is symmetric in all the indices. An interesting application in the Standard Model is to take  $t_a = Q$ , the matrix of electric charge, and  $t_b, t_c$  the isospin matrices. The constraint  $\text{tr} Q = 0$  is obeyed in the Standard Model since each generation of quark doublet of three colours is paired with a lepton doublet.

## 5 Global Chiral Anomaly and the Strong CP Problem.

Global U(1) axial anomaly is the sine qua non for the strong CP problem. The problem consists in the gross disagreement in the experimental data for the CP violating electric dipole moment of neutron (EDMN) which are consistent with a null result and theoretical estimates that invariably give a large value. Strong CP problem provides the unique arena where the concept of a global chiral anomaly is confronted with direct experimental data.

The two possible sources for CP violation in QCD action

$$S_{QCD} = S_G + \int \bar{q}(i\not{D})q d^4x + m \int \bar{q} e^{2i\alpha_{ew}\gamma_5} q d^4x + \theta_{QCD} \Delta S \quad \dots (5.1)$$

are the chiral phase  $\alpha_{ew}$  in the quark mass which arises from the electroweak sector of the Standard Model, and the QCD vacuum term with parameter  $\theta_{QCD}$

$$\theta_{QCD} \Delta S = \theta_{QCD} \frac{g^2}{32\pi^2} \epsilon_{\mu\nu\lambda\rho} \int \text{tr} F_{\mu\nu} F_{\lambda\rho} d^4x \quad \dots (5.2)$$

In (5.1)  $S_G$  represents the contributions from the gauge fields. For gauge fields with nontrivial topology the coefficient of  $\theta_{QCD}$  in (5.2) gives precisely the winding number  $\nu \neq 0$ ,

$$\nu = \frac{g^2}{32\pi^2} \epsilon_{\mu\nu\lambda\rho} \int \text{tr} F_{\mu\nu} F_{\lambda\rho} d^4x \quad \dots (5.3)$$

The chiral phase in the mass term in (5.1) can be transformed away by relabelling the quark fields

$$q \rightarrow e^{-i\alpha_{ew}\gamma_5} q, \quad \bar{q} \rightarrow \bar{q} e^{-i\alpha_{ew}\gamma_5} \quad \dots (5.4)$$

There relabelling, however, introduces a Jacobian

$$J(\alpha_{ew}) = \exp \left[ -i\alpha_{ew} \frac{g^2}{16\pi^2} \epsilon_{\mu\nu\lambda\rho} \int \text{tr} F_{\mu\nu} F_{\lambda\rho} d^4x \right] \quad \dots (5.5)$$

where the coefficient of  $\alpha_{ew}$  in the exponent is  $2\nu$ , i.e. twice the winding number of the gauge field configuration, which is nontrivial precisely in sectors where instantons live. The relabelling, therefore, merely shifts  $\alpha_{ew}$  to  $\theta_{QCD}$  giving an effective  $\bar{\theta}$

$$\bar{\theta} = \theta_{QCD} - 2N_f \alpha_{ew} \quad \dots (5.6)$$

where  $N_f$  is the number of quark flavours. All physical quantities in this scenario, therefore, depends on  $\bar{\theta}$  and not on  $\theta_{QCD}$  or  $\alpha_{ew}$  individually. Theoretical estimates<sup>22</sup> for CP-violating EDMN are all in the range

$$d_n^{th} \approx \bar{\theta} \times 10^{-15 \pm 1} e.cm \quad \dots (5.7)$$

Experimental data  $d_n^{ex} \leq 10^{-26} e.cm$ , therefore, suggests  $\bar{\theta} < 10^{-9}$ . Such a small value requires near cancellation of two parameters  $\theta_{QCD}$  and  $\alpha_{ew}$  as in (5.6), which arise from completely different sectors of the Standard Model. This is the strong CP problem, which is essentially a problem of fine tuning.

Attempts to remedy the strong CP problem by invoking a spontaneously broken global chiral U(1) symmetry, the Peccei-Quinn symmetry, have been pursued vigorously<sup>22</sup>. The idea essentially is that the effective  $\bar{\theta}$  becomes a dynamical variable in this scenario involving the field of the pseudoscalar Goldstone boson associated with the broken Peccei-Quinn symmetry. The dynamical  $\bar{\theta}$  could then settle down to a minimum consistent with the conservation of P and CP. The axion has been virtually ruled out by experiments and the strong CP problem in its original formulation is no closer to a resolution now than it was at the time of its conception<sup>22</sup>.

**Question of Global Chiral Anomaly :** In view of the prevailing impasse, with axion window virtually closed, it is worthwhile to reexamine critically the basic premises that lead up to the strong CP problem. The question of a nontrivial global chiral anomaly clearly stands out as the most vulnerable among these basic premises.

The chiral limit of the axial vector Ward identity in a instanton-like background gauge field was given in (3.23)

$$\langle \partial_\mu (\bar{\psi} \gamma_5 \gamma_\mu \psi) \rangle_{m=0} = 2 \sum \epsilon_i \phi_{0i}^+ \phi_{0i} - \frac{g^2}{16\pi^2} \epsilon_{\mu\nu\lambda\rho} t_r F_{\mu\nu} F_{\lambda\rho} \quad \dots (3.23)$$

where the zero modes  $\phi_{0i}(x)$  are a fallout of the nontrivial winding number  $\nu$  of the gauge field. It is natural to identify the right hand side of (3.23) as the density of global chiral anomaly in an instanton-like background. Its space-time integral, the global chiral anomaly, vanishes by the Atiyah-Singer index theorem (3.19). This patently contradicts a nontrivial Jacobian as in (5.5), the cornerstone of the strong CP problem. The popular perception of a nontrivial global chiral anomaly and hence a nontrivial Jacobian (5.5) not only leads to the strong CP problem but is afflicted with contradictions in the chiral limit.

The source of these afflictions is easily traced to the popular identification of the partition function with the determinant of the Dirac operator

$$Z_f[A]_{\nu \neq 0} = \det(i\mathcal{D} + m) \quad \dots (5.8)$$

which is unphysical in the chiral limit because of zero modes. A key to the problem is provided by the theorem<sup>23</sup> which states that there are no wrong chirality zero modes of the Dirac operator  $\mathcal{D}$ , i.e., in the Atiyah-Singer index theorem (3.19) positive (negative) chirality zero modes  $n_+(n_-)$  are associated with positive (negative) winding number  $\nu$ . Thus

$$\begin{aligned} \dim \ker(D_R D_L) &= 0, & \nu &\geq 0 \\ \dim \ker(D_L D_R) &= 0, & \nu &\leq 0 \end{aligned} \quad \dots (5.9)$$

where  $D_L, D_R = D_L^\pm$  are the Weyl components of the Dirac operator  $\mathcal{D}$

$$\mathcal{D} = \begin{pmatrix} 0 & D_L \\ D_R & 0 \end{pmatrix} \quad \dots (5.10)$$



The theorem (5.9), therefore, assures that the partition functions defined as

$$\begin{aligned} Z_f[A]_{\nu \geq 0} &= \det(D_R D_L + m^2) \\ Z_f[A]_{\nu \leq 0} &= \det(D_L D_R + m^2) \end{aligned} \quad \dots (5.11)$$

in the respective gauge field sectors, are not afflicted, unlike (5.8), with zero modes and hence have smooth chiral limits. In the trivial sector  $\nu = 0$  the two representations coincide.

The representations in (5.11) require that instead of the Dirac basis  $\{\phi_n(x)\}$  of Sec.3 we use eigenfunction sets of Weyl operators  $D_R D_L$  and  $D_L D_R$  appropriate respectively for positive and negative  $\nu$ . Thus for  $\nu \geq 0$ , one writes

$$\phi_n(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} \frac{1}{\lambda_n} D_L \phi_{nL}(x) \\ \phi_{nL}(x) \end{pmatrix}, \quad \phi_{-n}(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} \frac{1}{\lambda_n} D_L \phi_{nL}(x) \\ -\phi_{nL}(x) \end{pmatrix} \quad \dots (5.12)$$

where  $\phi_{nL}(x)$  are orthonormal eigenfunctions of the positive definite hermitian operator  $D_R D_L$

$$D_R D_L \phi_{nL}(x) = \lambda_n^2 \phi_{nL}(x) \quad \dots (5.13)$$

The set  $\{\phi_{nL}(x)\}$  with  $\lambda_n^2 > 0$  provides a complete set of functions in the Weyl basis in  $\nu \geq 0$  sector. In the resulting axial vector Ward identity<sup>24</sup>

$$\langle \partial_\mu (\bar{\psi} \gamma_5 \gamma_\mu \psi) \rangle_{\nu > 0} = 2m \langle \bar{\psi} \gamma_5 \psi \rangle - \left\{ \frac{g^2}{16\pi^2} \epsilon_{\mu\nu\lambda\rho} \text{tr} F_{\mu\nu} F_{\lambda\rho} - \sum \phi_{0i}^+ \phi_{0i} \right\} \quad \dots (5.14)$$

the contribution from the mass term on the right hand side now vanishes smoothly in the chiral limit. The global chiral anomaly given by the space-time integral of the chiral limit of the four divergence of the axial vector current, therefore, vanishes and instead of (5.5), we now have

$$\begin{aligned} J(\alpha_{ew})_{\nu \neq 0} &= \exp \left[ -i\alpha_{ew} \int \left\{ \frac{g^2}{16\pi^2} \epsilon_{\mu\nu\lambda\rho} \text{tr} F_{\mu\nu} F_{\lambda\rho} - \sum \epsilon_i \phi_{0i}^+ \phi_{0i} \right\} d^4 x \right] \\ &= 1 \end{aligned} \quad \dots (5.15)$$

The vanishing of the global chiral anomaly means that the chiral phase  $\alpha_{ew}$  in the quark mass in (5.1) is unphysical and can be transformed away trivially by a global chiral rotation (5.4) without affecting in any way  $\theta_{QCD}$ . The vacuum parameter  $\theta_{QCD}$  remains invariant. The crux of the strong CP problem, the problem of fine tuning, therefore, melts away. CP symmetry is ensured simply through the natural choice  $\theta_{QCD} = 0$ .

## 6 Concluding Remarks

Ever since its conception in the context of the problem of neutral pion decay into two photons, chiral anomaly has been a topic of abiding interest and challenge in particle physics. The interest stems in a large measure from the need to couple fermions chirally to gauge fields in building models in particle physics. The challenge consists in formulating a consistent and systematic regularisation scheme in chiral gauge theories.

The paper highlights and elucidates the seminal role of the mass term in the axial vector Ward identity in generating the local ABJ anomaly and the global U(1) axial anomaly. Gauge invariance demands that the fermion gets decoupled from the divergence of the U(1) axial vector current if it is very heavy. This identifies the ABJ anomaly with the asymptotic limit of the fermion mass term with sign reversed. On the other hand, the chiral limit ( $m = 0$ ) of the same mass term does not vanish and consists of contributions from fermion zero modes when the background gauge field has a nontrivial topology  $\nu \neq 0$ . The space time integral of the chiral limit cancels the integral of the ABJ anomaly, the (sign-reversed) asymptotic limit of the mass term, thanks to the Atiyah-Singer index theorem. This suggests, contrary to popular perception, that the Jacobian for global U(1)

chiral transformation is trivial even in an instanton background. The triviality of the Jacobian is realised in a representation of the fermion partition function in the Weyl basis (5.11) which has a null kernel space. The point of interest in all this is that there is no strong CP problem in an axionless physical world.

Current interest in lattice formulation of chiral gauge theory centres around Dirac operators for lattice fermion which obey the Ginsparg-Wilson<sup>16</sup> relation (4.4). Apart from redefining chiral symmetry on lattice, the Ginsparg-Wilson relation introduces nonlocality<sup>18</sup>. It is interesting to note that in continuum formulation also it is possible to define a gauge invariant partition function (4.8) but only at the cost of locality. Fermion averaging of the gauge current implemented with this partition function yields the covariant current. The consistent current which obeys integrability, can be generated with the covariant current as input. The covariant derivative of the consistent current thus obtained yields the minimal anomaly which obeys the Wess-Zumino consistency condition. Both the anomalies, covariant and consistent, and the distinction between the two currents vanish if the fermion belongs to anomaly free representation (4.26).

### Acknowledgement

It is a pleasure to acknowledge indebtedness to my collaborators Rabin Banerjee, Asit De, and Partha Mitra. I should also like to thank Amitabha Lahiri for discussions, and Sugata Mukherjee and S.K. Singh for help in preparing the manuscript.

### References

- [1] J. Steinberger, *Phys. Rev.* **76**, 1180 (1949); H. Fukuda and Y. Miyamoto, *Prog. Theor. Phys.* **4**, 347 (1949).
- [2] R. Jackiw, in *current Algebra and Anomalies*, Eds. S.B. Treiman et al (World Scientific, Singapore, 1985), p.81.
- [3] D.G. Sutherland, *Nucl. Phys.* **B2**, 433 (1967); M. Veltman, *Proc. Roy. Soc.* **A301**, 107 (1967).
- [4] S. Adler, *Phys. Rev.* **177**, 2426 (1969); J.S. Bell and R. Jackiw, *Nuovo Cimento* **60A**, 47 (1969).
- [5] T. Appelquist and J. Carazzone, *Phys. Rev.* **D11**, 2856 (1975).
- [6] S. Adler, in *Lectures on Elementary Particles and Quantum Field Theory*, proceedings of 1970 Brandeis University Summer Institute in Theoretical Physics, Vol.1, Edited by Stanley Deser et al (M.I.T. Press, Cambridge, Mass, 1970).
- [7] L. Rosenberg, *Phys. Rev.* **129**, 2786 (1963).
- [8] S. Adler and W.A. Bardeen, *Phys. Rev.* **182**, 157 (1969).
- [9] K. Fujikawa, *Phys. Rev. Lett.* **42**, 1195 (1979); *Phys. Rev.* **D21**, 2848 (1980).
- [10] K. Osterwalder and R. Schrader, *Helv. Phys. Acta.* **46**, 277 (1973).
- [11] M. Mehta, *Phys. Rev. Lett.* **65**, 1983 (1990); P. van Nieuwenhuizen and A. Waldron, *Phys. Lett* **B389**, 29 (1996).
- [12] H. Banerjee, P. Mitra and D. Chatterjee, *Z. Phys.* **C62**, 511 (1994); H. Banerjee, *Ind. J. of Phys. (Spl.)* **80**, 333 (1997).
- [13] M. Atiyah, R. Bott and V. Patodi, *Invent. Math.* **19**, 279 (1973).

- [14] H. B. Nielsen and M. Ninomiya, *Phys. Lett.* **B105**, 219 (1981); *Nucl. Phys.* **B185**, 20 (1981).
- [15] K.G. Wilson, in *new Phenomena in Subnuclear Physics* (Erice, 1975), ed. A. Zichichi (Plenum Press, New York, 1977), p.69.
- [16] P.H. Ginsparg and K.G. Wilson, *Phys. Rev.* **D25**, 2649 (1982).
- [17] M. Luscher, *Phys. Lett.* **B428**, 342 (1998).
- [18] I. Hovarth, *Phys. Rev. Lett.* **81**, 4063 (1998).
- [19] H. Banerjee and R. Banerjee, *Phys. Lett.* **B174**, 313 (1986); H. Banerjee, R. Banerjee and P. Mitra, *Z. Phys.* **C32**, 445 (1986).
- [20] W.A. Bardeen and B. Zumino, *Nucl. Phys.* **B244**, 421 (1984); H. Leutwyler, *Phys. Lett.* **B152**, 78 (1985).
- [21] J. Wess and B. Zumino, *Phys. Lett.* **B37**, 95 (1971).
- [22] R.D. Peccei, in *CP Violation*, ed. C. Jarlskog (World Scientific, Singapore, 1989), p.503.
- [23] S. Coleman, in *Aspects of Symmetry* (Cambridge University Press, Cambridge, 1985) p.265.

# 11. Coherent States in Field Theory

Wei-Min Zhang \*

Department of Physics, National Cheng-Kung University,  
Tainan, Taiwan 701, R.O.C.

## Abstract

Coherent states have three main properties: coherence, overcompleteness and intrinsic geometrization. These unique properties play fundamental roles in field theory, especially, in the description of classical domains and quantum fluctuations of physical fields, in the calculations of physical processes involving infinite number of virtual particles, in the derivation of functional integrals and various effective field theories, also in the determination of long-range orders and collective excitations, and finally in the exploration of origins of topologically nontrivial gauge fields and associated gauge degrees of freedom.

## 1 Introduction

In the past thirty-six years, the developments and applications of coherent states have been made tremendous progress. Yet, the idea of creating a coherent state for a quantum system was conceived well before that. In fact, back in 1926, Schrödinger first proposed the idea of what is now called “coherent states” [1] in connection with the quantum states of classical motion for a harmonic oscillator. In other words, the coherent states were invented immediately after the birth of quantum mechanics. However, between 1926 and 1962, activities in this field remained almost dormant, except for a few works in condensed matter physics [2, 3, 4] and particle physics [5, 6] in 50’s. It was not until some thirty five years after Schrödinger’s pioneering paper that the first modern and systematic application to field theory was made by Glauber and Sudarshan [7, 8] and launched this fruitful and important field of study in theoretical as well as experimental physics.

I became interested in the subject of coherent states about fifteen years ago. On the occasion of Prof. Sudarshan visiting Suzhou of China (1984), I listened for the first time in life a topic on coherent states presented by Prof. Sudarshan. As a second-year graduate student at that time, I was looking for some research problem on collective excitations in strongly interacted many-body systems (particularly in nuclear physics). Prof. Sudarshan’s lecture inspired me to think whether under constraint(s) of dynamical symmetries collective excitations can be described in terms of coherent states, as a result of multi-particle correlations (coherence). Later on I realized, this is indeed a very active subject covering problems from condensed matter physics to nuclear and particle physics. Of course, these coherent states have no longer the simple but beautiful form Glauber and Sudarshan proposed for light beams. Actually, these states are generated by complicated collective composite operators of particle-particle pairs or particle-hole pairs. Their mathematical structure were already developed in early 70’s by Perelomov and Gilmore [9, 10] based on the theory of Lie groups. Nowadays, the concept of coherent states has been extensively investigated. Many methods based on coherent states have also been developed for various theoretical problems. Nevertheless, the original development of coherent states in quantum electromagnetic field (or more precisely, in the study of quantum optical coherence) has made tremendous influence in physics.

One can find that a large body of the literature on coherent states has appeared. This vast literature was exhaustively collected, catalogued and classified by Klauder and Skagerstam [11]. About the mathematical usefulness of coherent states as a new tool to study the unitary representations of

---

\*E-mail: wzhang@mail.ncku.edu.tw

Lie groups has been described in a well expository book by Perelomov [12]. A review article on the theory of coherent states and its applications that cover subjects of quantum mechanics, statistical mechanics, nonlinear dynamics and many-body physics has also been presented by author and his collaborators [13]. In this article, I will only concentrate on the topic of coherent states in field theory. As usual, it is not my intention to give a complete review about coherent states in field theory. An extensive review on coherent states in field theory and particle physics may be found in [14]. I will rather like to present here a discussion on whether one can formulate field theory in terms of coherent states such that the new formulation may bring some new insights to the next development of field theory in the new millennium. Coherent state can become a useful and important subject in physics because of its three unique properties: the coherence, the overcompleteness and the intrinsic geometrization. These unique properties, in certain contents, are fundamental to field theory. I will select some typical topics in field theory that can be efficiently described by coherent states based on these properties. These topics include the productions of coherent states in field theory, the basic formulation of quantum field theory in terms of coherent state functional integrals, the spontaneously symmetry breaking described from coherent states, and the effective field theories derived from coherent states. Also, I will “sprinkle” discussions about the geometrical phases of coherent states and their interpretation as gauge degrees of freedom in field theory, a subject which has still received increasing importance in one’s attempt to understand the fundamental of nature.

## 2 Photon coherent states

I may begin with the simplest coherent state of photons, or more generally speaking, bosons. Such a set of coherent states has been described in most of quantum mechanics text books and are familiar to most of physicists. It is indeed the most popular coherent state that has been used widely in various fields. The coherent state of photons can describe not only the coherence of electromagnetic field, but also many other properties of bosonic fields. It is the basis of modern quantum optics [15], and it also provides a fundamental framework to quantum field theory, as one will see later.

By means of **optical coherence**, one may consider the  $n$ -th order correlation function of electromagnetic field:

$$G^n(x_1, \dots, x_n, x_{n+1}, \dots, x_{2n}) = \text{tr}\{\rho E^-(x_1) \cdots E^-(x_n) E^+(x_{n+1}) \cdots E^+(x_{2n})\}, \quad (1)$$

where  $x_i$  is the time-space coordinates,  $\rho$  denotes the density operator, and  $E^\pm(x_i)$  represent the electric field operators with positive and negative frequency. For simplification, the polarization of electric field is fixed. According to Glauber [7] the complete coherence of a radiation field is that all of the correlation functions satisfy the following factorization condition:

$$G^n(x_1, \dots, x_n, x_{n+1}, \dots, x_{2n}) = \mathcal{E}^*(x_1) \cdots \mathcal{E}^*(x_n) \mathcal{E}(x_{n+1}) \cdots \mathcal{E}(x_{2n}). \quad (2)$$

This condition implies electric field operators must behave like classical field variables. It may also indicate the electric field operator should have its own eigenstates with the corresponding classical field variables as its eigenvalues:

$$E^+(x_i)|\phi\rangle = \mathcal{E}(x_i)|\phi\rangle, \quad \langle\phi|E^-(x_i) = \langle\phi|\mathcal{E}^*(x_i). \quad (3)$$

Moreover, the density operator must also be expressed in terms of the eigenstates  $|\phi\rangle$ . Obviously, the conventional Fock space in quantum theory does not obey the above condition.

This is actually a nontrivial problem, because it requires a complete description of classical motions in terms of quantum states. Meantime, the operator  $E^\pm(x_i)$  itself is not a Hermitian operator. The eigenstate problem of a nonhermitian operator is unusual in quantum mechanics. Fortunately, such quantum states have already been constructed by Schrödinger soon after his invention of quantum mechanics in 1926. In order to answer the question how microscopic dynamics transits to macroscopic world, Schrödinger looked for quantum states which follow precisely the

corresponding classical trajectories all the time, and meantime, the states must also be the exact solution of quantum dynamical equation (i.e., the Schrödinger equation). But only for harmonic oscillator, such states were constructed [1]:

$$\phi_z(x) \sim \exp \left\{ -\frac{1}{2}(x+z)^2 \right\}, \quad (4)$$

where  $z$  is a complex variable. These states are actually the Gaussian wave packets centered on the classical trajectory  $z = (x + ip)$ ,  $x$  and  $p$  are the position of harmonic oscillator in the phase space that satisfies classical equations of motion. One can show that Eq. (4) is also an exact solution of Schrödinger equation. The classicality of Gaussian wave packets are manifested by the minimum uncertainty relationship:

$$\Delta x^2 \Delta p^2 = \frac{\hbar^2}{4} \quad \text{and} \quad \Delta p = \Delta x. \quad (5)$$

In other words, the wave packets governed by the Hamiltonian of harmonic oscillator follow classical trajectories and do not spread in time.

Glauber and Sudarshan discovered [7, 8] that such a wave packet is a superposition of Fock states. It is also an eigenstate of  $E^+(x_i)$ . In quantum field theory, electromagnetic field consists of infinite harmonic oscillating modes (photons). Explicitly, the Hamiltonian of quantum electromagnetic field (in Coulomb gauge) is given by

$$H = -\frac{1}{2} \int d^3x \{ \mathbf{E}^2 + \mathbf{B}^2 \}, \quad (6)$$

where  $\mathbf{E}$  and  $\mathbf{B}$  are the electric and magnetic fields. The electromagnetic field can be expressed by the vector potential  $\mathbf{A}$ :  $\mathbf{E} = -\partial \mathbf{A} / \partial t$ ,  $\mathbf{B} = \nabla \times \mathbf{A}$ . It is convenient to expand the vector potential in terms of plane waves (Fourier series)

$$\mathbf{A}(x, t) = \int \frac{d^3k}{\sqrt{(2\pi)^3 2\omega_k}} \sum_{\lambda} \left\{ a_k^{\lambda} \bar{\varepsilon}^{\lambda}(k) e^{-ikx} + a_k^{\lambda\dagger} \bar{\varepsilon}^{\lambda*}(k) e^{ikx} \right\}, \quad (7)$$

where  $\bar{\varepsilon}^{\lambda}(k)$  is the polarization vector of electromagnetic field, and  $(a_k^{\lambda\dagger}, a_k^{\lambda})$  are the creation and annihilation operators,

$$[a_k^{\lambda}, a_{k'}^{\lambda'\dagger}] = \delta_{\lambda\lambda'} \delta_{kk'}, \quad [a_k^{\lambda}, a_{k'}^{\lambda'}] = [a_k^{\lambda\dagger}, a_{k'}^{\lambda'\dagger}] = 0. \quad (8)$$

Then the Hamiltonian of electromagnetic field can be deduced to

$$H = \sum_{k\lambda} \omega_k (a_k^{\lambda\dagger} a_k^{\lambda} + 1/2), \quad (9)$$

which means that the electromagnetic field consists of infinite individual electromagnetic modes, i.e., photons. Each photon corresponds to a harmonic oscillator.

In the particle number representation, the Gaussian wave packet can be written as

$$|z\rangle = \exp\left(-\frac{1}{2}|z|^2\right) \exp(za^{\dagger})|0\rangle. \quad (10)$$

where  $|0\rangle$  is the vacuum state:  $a|0\rangle = 0$ . From the above expression, it is easy to show that the wave packet is also an eigenstate of the annihilation operator  $a$ :

$$a|z\rangle = z|z\rangle, \quad (11)$$

Thus, the quantum state describing the optical coherence of electromagnetic field can be expressed by

$$|\{z_k^{\lambda}\}\rangle = \exp \left\{ -\frac{1}{2} \int d^3k \sum_{\lambda} |z_k^{\lambda}|^2 \right\} \exp \left\{ \int d^3k \sum_{\lambda} z_k^{\lambda} a_k^{\lambda\dagger} \right\} |0\rangle, \quad (12)$$

which is an eigenstate of the positive frequency part of the electric field operator,

$$\mathbf{E}^+(x)|\{z_k^\lambda\}\rangle = \mathcal{E}(x)|\{z_k^\lambda\}\rangle, \quad \mathcal{E}(x) = i \int \frac{d^3k}{(2\pi)^{3/2}} \sqrt{\frac{\omega_k}{2}} \sum_\lambda z_k^\lambda \vec{\epsilon}_\lambda(k) e^{-i(\omega_k t - \mathbf{k} \cdot \mathbf{x})}. \quad (13)$$

Besides, the above state has another very important property: it supports the following resolution of identity:

$$\int |\{z_k^\lambda\}\rangle \langle \{z_k^\lambda\}| \prod_{k\lambda} \frac{dz_k^\lambda dz_k^{\lambda*}}{2\pi} = I. \quad (14)$$

In other words, these states in the complex space (in terms of the variable  $z$ ) form a complete set of states (more precisely speaking, it is overcomplete because of the continuity of these states). This complete set is certainly very different from the set of Fock states. Because of the overcompleteness and the analyticity of these states, one can expand the density operator by (12) in a diagonal form (the so-called P-representation [7]):

$$\begin{aligned} \rho &= \int P(\{z_k^\lambda\}) |\{z_k^\lambda\}\rangle \langle \{z_k^\lambda\}| \prod_{k\lambda} dz_k^\lambda dz_k^{\lambda*}, \\ \text{tr} \rho &= \int P(\{z_k^\lambda\}) \prod_{k\lambda} dz_k^\lambda dz_k^{\lambda*} = 1. \end{aligned} \quad (15)$$

where  $P(z)$  is a weight function. In terms of these states (12), the factorization criterion of coherent light beams is automatically satisfied. Glauber named such states the *coherent states*. To be more specific, one may call them the “photon coherent states”. Physically, the photon coherent states have a well-defined phase for each mode. Therefore, coherent light beams can be completely described in quantum mechanics in terms of photon coherent states. For those who wish to have more detailed discussion on physical consequences of the photon coherent states in quantum optics, please refer to the excellent book by Klauder and Sudarshan [15].

### 3 Coherent states and $S$ -Matrix

As we have seen, the photon coherent state was introduced by the requirement of optical coherence. Here, I may ask a more general question, namely, how are photon coherent states generated in field theory? In field theory, all physical quantities are derivable from the vacuum-to-vacuum transition amplitude in the presence of external sources. It can show that the final state in such processes is a coherent state if there is no other interactions except for a linear interaction with the external field.

To be specific, one may consider the electromagnetic field interacting with a classical source:

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} - A_\mu j^\mu, \quad (16)$$

where the classical source is a conserved current:  $\partial^\mu j_\mu = 0$ . In the Feynman gauge, the equation of motion for the electromagnetic field is given by

$$\partial_\mu F^{\mu\nu} = \partial_\mu^2 A^\nu = j^\nu. \quad (17)$$

A general solution of the above equation is

$$A^\mu(x) = A_0^\mu(x) + \int d^4y \Delta(x-y) j^\mu(y), \quad (18)$$

where  $A_0^\mu(x)$  is the solution of free field, and  $\Delta(x-y)$  is the Green function determined by  $\partial_x^2 \Delta(x-y) = \delta^4(x-y)$ . If one assumes that the interaction is switched on adiabatically in a

finite time interval, then

$$\begin{aligned} A^\mu(x) &= A_{\text{in}}^\mu(x) + \int d^4y \Delta_{\text{ret}}(x-y) j^\mu(y) \\ &= A_{\text{out}}^\mu(x) + \int d^4y \Delta_{\text{adv}}(x-y) j^\mu(y), \end{aligned} \quad (19)$$

where the retarded and advanced Green functions are given by

$$\Delta_{\text{adv}}^{\text{ret}}(x) = -\frac{1}{(2\pi)^4} \int d^4p \frac{e^{-ipx}}{(p_0 \pm i\epsilon)^2 - \mathbf{p}^2}, \quad (20)$$

and  $A_{\text{in}}^\mu$  and  $A_{\text{out}}^\mu$  are free fields describing the photon field before and after its interaction with the classical course  $j^\mu$ . The corresponding photon states are the in- and out-states (denoted by  $|\rangle_{\text{in}}$  and  $|\rangle_{\text{out}}$ , respectively). The in- and out-states form two complete sets of free states constructed as a Fock space by the free field operators  $A_{\text{in}}^\mu$  and  $A_{\text{out}}^\mu$ . Therefore, there must exist a unitary transformation  $S$  (namely  $S$ -matrix) to connect these two complete sets:

$$A_{\text{out}}^\mu = S^\dagger A_{\text{in}}^\mu S, \quad |\rangle_{\text{out}} = S^\dagger |\rangle_{\text{in}}. \quad (21)$$

From (19), one can see that

$$\begin{aligned} A_{\text{out}}^\mu(x) &= A_{\text{in}}^\mu(x) + \int d^4y [\Delta_{\text{ret}}(x-y) - \Delta_{\text{adv}}(x-y)] j_\mu(y) \\ &= A_{\text{in}}^\mu(x) + A_{\text{cl}}^\mu(x), \end{aligned} \quad (22)$$

and  $A_{\text{cl}}^\mu(x)$  is a  $c$ -number (classical) field generated by the classical current  $j^\mu(x)$ . Notice that  $\Delta_{\text{adv}}(x) - \Delta_{\text{ret}}(x) = \Delta(x)$  which relates to the commutator of free fields  $[A_{\text{in}}^\mu(x), A_{\text{in}}^\nu(y)] = [A_{\text{out}}^\mu(x), A_{\text{out}}^\nu(y)] = -ig^{\mu\nu} \Delta(x-y)$ . One may check that the  $S$ -matrix can be written as

$$S = \exp \left\{ -i \int d^4x A_{\text{in}} \cdot j(x) \right\} = \exp \left\{ -i \int d^4x A_{\text{out}} \cdot j(x) \right\}. \quad (23)$$

If we start at time  $-\infty$  from the vacuum state  $|0\rangle_{\text{in}}$ , the final state after the free field  $A^\mu(x)$  interacted with the classical current  $j^\mu(x)$  becomes a coherent state:

$$|0\rangle_{\text{out}} = \exp \left\{ i \int d^4x A_{\text{in}}(x) \cdot j(x) \right\} |0\rangle_{\text{in}}. \quad (24)$$

In terms of the Fourier expansion,

$$A_\mu(x) = \int \frac{d^3k}{\sqrt{2(2\pi)^2|\mathbf{k}|}} \sum_\lambda \left\{ a_k^\lambda \varepsilon_\mu^\lambda(k) e^{-ikx} + a_k^{\lambda\dagger} \varepsilon_\mu^{\lambda*}(k) e^{ikx} \right\}. \quad (25)$$

The final state can be expressed as

$$|0\rangle_{\text{out}} = \exp \left\{ -\frac{1}{2} \int d^3k \sum_\lambda |z_k^\lambda|^2 \right\} \exp \left\{ \int d^3k \sum_\lambda z_k^\lambda a_k^{\lambda\dagger} \right\} |0\rangle_{\text{in}} = |\{z_k^\lambda\}\rangle, \quad (26)$$

where  $z_k^\lambda = \varepsilon^\lambda(k) \cdot j(k)$ , and  $j^\mu(k)$  is the Fourier transform of the classical current  $j^\mu(x)$ . This is the same photon coherent states introduced by Glauber in the study of quantum optical coherence.

Indeed, one can derive similarly the photon coherent state for the laser beams (discussed in the previous section) from a more microscopic picture. The Hamiltonian in quantum optics that describes the interaction between  $N$  atoms and the electromagnetic field can be written as:

$$H = \sum_k \omega_k a_k^\dagger a_k + \sum_i \epsilon_i \sigma_0^i + \sum_{k,i} \gamma_{ki}(t) \left\{ \frac{\sigma_+^i}{\sqrt{N}} a_k + \frac{\sigma_-^i}{\sqrt{N}} a_k^\dagger \right\}, \quad (27)$$



where  $\gamma_{ki}$  are the coupling coefficients between atoms and electromagnetic field. One of the crucial assumptions made in the construction of the above Hamiltonian for laser beams is that each of the  $N$  atoms, labeled by the index  $i$ , is a two-level system and therefore its dynamical variables are the usual Pauli operators  $\{\sigma_0^i, \sigma_+^i, \sigma_-^i\}$ . Furthermore, the atomic variables can be treated as a classical source (i.e. the spin operators  $\sigma^i$  can be regarded as  $c$ -numbers), and the coupling strength  $\gamma_{ki}$  are identical for all the atoms (i.e.  $\gamma_{ki} = \gamma_k$ ). Then, Eq. (27) is reduced to

$$\begin{aligned} H &= \sum_k \omega_k a_k^\dagger a_k + \sum_i \epsilon \langle \sigma_0^i \rangle + \sum_{k,i} \gamma_{ki}(t) \left\{ \frac{\langle \sigma_+^i \rangle}{\sqrt{N}} a_k + \frac{\langle \sigma_-^i \rangle}{\sqrt{N}} a_k^\dagger \right\} \\ &= \sum_k \omega_k a_k^\dagger a_k + \sum_k \left[ \lambda_k(t) a_k^\dagger + \lambda_k^*(t) a_k \right] + \text{const.} \end{aligned} \quad (28)$$

This corresponds to the electromagnetic field interacting with an external time dependent source. The general solution of the Schrödinger equation for this Hamiltonian is the photon coherent states. This provides the microscopic picture how the photon coherent state is generated and why it becomes the fundamental of quantum optics.

Soon after the development of coherent states in optical coherence, it was found that the photon coherent state also plays an important role in solving the **infrared divergence** in quantum electrodynamics for electron scatterings [16, 17, 18, 19, 20] (also see the review by Papanicolaou [21]). As it is well known, the matrix element in quantum electrodynamics for the scattering of an initial state containing a finite number of electrons and photon into a similar final state has a logarithmical infrared divergence for the small momentum  $k$  [22]. This is because in an actual scattering experiment, electromagnetic fields interact with the source particles so that an infinite number of soft photons are emitted. These emitted soft photons form a coherent state to the final state, as we have discussed. To be more specific, consider a single electron scattering. The source particle can be represented by a classical current. The Fourier transform of the classical current is given by

$$j^\mu(k) = \frac{ie}{\sqrt{2(2\pi)^3|k|}} \left( \frac{p_f^\mu}{p_f \cdot k} - \frac{p_i^\mu}{p_i \cdot k} \right), \quad (29)$$

where  $p_{i,f}$  are the electron's momentum in the initial and final states. If one sums the cross sections over all possible final states containing any number of soft photons with momenta below the threshold of observability [by using the photon coherent state (26) with the above classical source], the infrared divergence is canceled. This gives a beautiful solution to the infrared problem in quantum electrodynamics.

Moreover, one can also show that if the matrix element for scatterings are calculated with the initial and final states containing infinite number of soft photons by the photon coherent state, the infrared divergences are canceled order by order at matrix element level (not only in cross sections) [16]. The photon coherent state may also use to remove the similar infrared problem in quantum gravity, as noticed by Weinberg [23]. These are perhaps the second important applications of the photon coherent state in field theory. In addition, one has also attempted to use coherent states to treat infrared divergences in non-abelian gauge theory [24]. However, in the non-abelian gauge theory, the infrared divergence is much more complicated [25, 26]. It contains two-type infrared divergences, the massless soft infrared divergence and collinear divergence. It is not clear whether one can construct some non-abelian coherent states to handle both the soft and collinear infrared divergences.

## 4 Functional integrals in field theory

When the quantum fields interact with quantum fields rather than classical external fields, the  $S$ -matrix (or the time-evolution operator) does not generate coherent states from the incoming vacuum. In such cases, coherent states are useful in the derivation of functional integral in field theory. Quantum field theory can be reformulated in terms of coherent states not only because

of its classicality and being eigenstates of the annihilation operator. As we have mentioned, the coherent states are overcomplete:

$$\int |z\rangle\langle z| \frac{dz dz^*}{2\pi} = I. \quad (30)$$

All these three properties (the classicality, the eigenstates of the positive frequency part of field operator and the overcompleteness) together allow one to reformulate quantum field theory in terms of a functional integral. Actually, the content of this section can be found in many text books, but for completeness I will repeat these discussions here.

The ordinary path integral of quantum mechanics developed by Feynman [27] can be obtained from the evolution operator by writing the evolution operator as

$$U(t_f, t_0) = \exp \left\{ -iH(t_f - t_0) \right\} = \lim_{N \rightarrow \infty} \left( \exp \left\{ iH \frac{t_f - t_0}{N} \right\} \right)^N \quad (31)$$

and then inserting a resolution of the identity in terms of the position states

$$\int dx |x\rangle\langle x| = I \quad (32)$$

between the terms of above product. This results in the familiar path integral of quantum mechanics,

$$\begin{aligned} \langle x'(t_f) | x(t_0) \rangle &= \langle x' | U(t_f, t_0) | x \rangle \\ &= \int [dx(t)] \exp \left\{ i \int_{t_0}^{t_f} dt \mathcal{L}(x(t), \dot{x}(t)) \right\}, \end{aligned} \quad (33)$$

where  $\mathcal{L}$  is a classical Lagrangian which generally has a form of

$$\mathcal{L}(x, \dot{x}) = \frac{1}{2} \left( \frac{dx}{dt} \right)^2 - V(x), \quad (34)$$

and  $[dx(t)] \equiv \prod_{t_0 \leq t \leq t_f} dx(t)$  is a functional measure of the path integration [28].

Instead of using the basis of the position eigenstates (32), we may use the coherent state basis and insert the resolution of identity (30) between the terms of product (31). Then a phase space formulation of path integrals can be obtained as was first proposed by Klauder [6, 30] (More detailed derivation will be given later in the application to field theory),

$$\begin{aligned} \langle z'(t_f) | z(t_0) \rangle &= \langle z' | U(t_f, t_0) | z \rangle \\ &= \int [dx(t)] \left[ \frac{dp(t)}{2\pi} \right] \exp \left\{ i \int_{t_0}^{t_f} dt \mathcal{L}(x(t), p(t)) \right\}, \end{aligned} \quad (35)$$

with

$$\begin{aligned} \mathcal{L}(x, p) &= \langle z | i \frac{d}{dt} | z \rangle - \langle z | H | z \rangle \\ &= \frac{1}{2} \left( p \frac{dx}{dt} - x \frac{dp}{dt} \right) - \mathcal{H}(x, p). \end{aligned} \quad (36)$$

where  $z = (x + ip)/\sqrt{2}$  and  $z^* = (x - ip)/\sqrt{2}$ , with the initial and final positions  $x(t_0)$  and  $x(t_f)$  fixed. This derivation of Feynman's path integral is particularly useful in obtaining a functional integral of quantum field theory.

To derive a functional integral of quantum field theory, we may start with a neutral scalar field  $\phi(x)$ , for simplification. The Lagrangian density of a neutral field is given by

$$\mathcal{L} = \frac{1}{2} [(\partial\phi)^2 - m^2\phi^2] - V(\phi), \quad (37)$$

where  $V(\phi)$  represents a self-interacting potential, such as  $\frac{\lambda}{4!}\phi^4(x)$ . The canonical momentum density conjugate to  $\phi(\mathbf{x}, t)$  is determined by  $\pi(\mathbf{x}, t) = \partial\mathcal{L}/\partial\dot{\phi}(\mathbf{x}, t)$ . Then the canonical quantization leads to

$$[\phi_{op}(\mathbf{x}, t), \pi_{op}(\mathbf{x}', t)] = i\delta^3(\mathbf{x} - \mathbf{x}'), \quad (38)$$

In the plane-wave expansion, one has

$$\begin{aligned} \phi_{op}(\mathbf{x}, t) &= \int \frac{d^3k}{(2\pi)^{3/2}} \sqrt{\frac{1}{2\omega_k}} \{a_k e^{-ikx} + a_k^\dagger e^{ikx}\}, \\ \pi_{op}(\mathbf{x}, t) &= -i \int \frac{d^3k}{(2\pi)^{3/2}} \sqrt{\frac{\omega_k}{2}} \{a_k e^{-ikx} - a_k^\dagger e^{ikx}\}, \end{aligned} \quad (39)$$

and the quantum Hamiltonian can be written as

$$H(t) = \int d^3\mathbf{x} : \left\{ \frac{1}{2}[\pi_{op}^2 + (\nabla\phi_{op})^2 + m^2\phi_{op}^2] + V(\phi_{op}) \right\} : \quad (40)$$

here  $:$  denotes the normal ordering with respect to the creation and annihilation operators  $a_k^\dagger$  and  $a_k$ . Now, one can define the scalar field coherent state as

$$\begin{aligned} |\phi\pi\rangle &= \exp \left\{ i \int d^3x [\pi(x)\phi_{op}(x) - \phi(x)\pi_{op}(x)] \right\} |0\rangle \\ &= \exp \left\{ -\frac{1}{2} \int d^3k |z_k|^2 \right\} \exp \left\{ \int d^3k (z_k a_k^\dagger) \right\} |0\rangle, \end{aligned} \quad (41)$$

from which a functional integral of field theory can be derived explicitly. Note that the coherent state in field theory is defined at a given instant time  $t$  over the whole space  $\{\mathbf{x}\}$ .

Since the Hamiltonian formalism of field theory is the same as in quantum mechanics, one can directly calculate the time Green's function defined as the matrix element of the evolution operator in coherent state basis,

$$G(t_f, t_0) = \langle \phi' \pi' | U(t_f, t_0) | \phi \pi \rangle = \langle \phi' \pi' | T \exp \left\{ -i \int_{t_0}^{t_f} dt H(t) \right\} | \phi \pi \rangle, \quad (42)$$

where  $T$  is the time-ordering operator. One may slice the time interval  $t_f - t_0$  into  $N$  equal segments:  $\varepsilon = (t_f - t_0)/N$  so that in the sense of  $N \rightarrow \infty$ , the evolution operator can be written as a subsequently multiplication of the evolution operator in the interval  $\varepsilon$ :

$$\begin{aligned} U(t_f, t_0) &= \exp \left\{ -i\varepsilon H(t_n) \right\} \exp \left\{ -i\varepsilon H(t_{n-1}) \right\} \\ &\quad \cdots \exp \left\{ -i\varepsilon H(t_2) \right\} \exp \left\{ -i\varepsilon H(t_1) \right\}. \end{aligned} \quad (43)$$

Using the same procedure as in the derivation of Feynman's path integral in quantum mechanics, one should insert the resolution of identity,

$$\int [d\phi(\mathbf{x})] \left[ \frac{d\pi(\mathbf{x})}{2\pi} \right] |\phi\pi\rangle \langle \phi\pi| = I, \quad (44)$$

at each interval point, where  $[d\phi(\mathbf{x})] \equiv \prod_{-\infty < \mathbf{x} < \infty} d\phi(\mathbf{x})$ , etc. are defined over the whole space. Then

$$G(t_f, t_0) = \lim_{N \rightarrow \infty} \int \left( \prod_{i=1}^{N-1} [d\phi_i(\mathbf{x})] \left[ \frac{d\pi_i(\mathbf{x})}{2\pi} \right] \right) \prod_{i=1}^N \langle \phi_i \pi_i | \exp \left\{ -i\varepsilon H(t_i) \right\} | \phi_{i-1} \pi_{i-1} \rangle. \quad (45)$$

Up to the first order in  $\varepsilon$ ,

$$\langle \phi_i \pi_i | \exp(-i\varepsilon H(t_i)) | \phi_{i-1} \pi_{i-1} \rangle \approx \langle \phi_i \pi_i | \phi_{i-1} \pi_{i-1} \rangle \exp \left( -i\varepsilon \frac{\langle \phi_i \pi_i | H(t_i) | \phi_{i-1} \pi_{i-1} \rangle}{\langle \phi_i \pi_i | \phi_{i-1} \pi_{i-1} \rangle} \right). \quad (46)$$

Note that the coherent state  $|\phi\pi\rangle$  is normalized. In the limit of  $\varepsilon \rightarrow 0$  (i.e.  $N \rightarrow \infty$ ),

$$\begin{aligned}\langle\phi_i\pi_i|\phi_{i-1}\pi_{i-1}\rangle &= 1 - \langle\phi_i\pi_i|(|\phi_i\pi_i\rangle - |\phi_{i-1}\pi_{i-1}\rangle)\rangle \\ &\simeq \exp\left\{i\varepsilon\langle\phi_i\pi_i|i\frac{\Delta|\phi_i\pi_i\rangle}{\varepsilon}\right\},\end{aligned}\quad (47)$$

where  $\Delta|\phi_i\pi_i\rangle \equiv |\phi_i\pi_i\rangle - |\phi_{i-1}\pi_{i-1}\rangle$ . Then, the Green's function becomes

$$\begin{aligned}G(t_f, t_0) &= \lim_{N \rightarrow \infty} \int \left( \prod_{i=1}^{N-1} [d\phi_i(x)] \left[ \frac{d\pi_i(x)}{2\pi} \right] \right) \exp i \sum_{i=1}^N \varepsilon \left\{ \langle\phi_i\pi_i|i\frac{\Delta|\phi_i\pi_i\rangle}{\varepsilon} - \langle\phi_i\pi_i|H(t_i)|\phi_i\pi_i\rangle \right\} \\ &= \int [d\phi(x)] \left[ \frac{d\pi(x)}{2\pi} \right] \exp \left\{ i \int_{t_0}^{t_f} dt \left[ \langle\phi\pi|i\frac{d}{dt}|\phi\pi\rangle - \langle\phi\pi|H|\phi\pi\rangle \right] \right\} \\ &= \int [d\phi(x)] \left[ \frac{d\pi(x)}{2\pi} \right] \exp \left\{ i \int_{t_0}^{t_f} dt \int d^3x \left[ \frac{1}{2}(\pi\dot{\phi} - \dot{\phi}\pi) - \mathcal{H}(x) \right] \right\},\end{aligned}\quad (48)$$

with

$$\mathcal{H} = \frac{1}{2}[\pi^2 + (\nabla\phi)^2 + m^2\phi^2] + V(\phi). \quad (49)$$

As we see that the coherent state gives a natural derivation of path integrals in field theory.

In field theory, the correlations between  $n$  fields are defined by the  $n$ -point Green functions,

$$G^{(n)}(x_1, \dots, x_n) = \langle 0|T(\phi(x_1) \cdots \phi(x_n))|0\rangle, \quad (50)$$

which can be determined from the generating functional  $W(J)$  which is defined as the vacuum-to-vacuum amplitude in the presence of external current  $J(x)$ :

$$W(J) = \langle 0|U(-\infty, \infty)|0\rangle_J. \quad (51)$$

This generating functional can then be expressed in terms of the time Green's function  $G(t_f, t_0)$  by adding a term  $\int d^3x J(x)\phi(x)$  in the exponent and then taking  $t_0 \rightarrow -\infty$  and  $t_f \rightarrow \infty$ :

$$\begin{aligned}W(J) &= \int [d\phi(x)] \left[ \frac{d\pi(x)}{2\pi} \right] \exp \left\{ \int d^4x \left[ \frac{1}{2}(\pi\dot{\phi} - \dot{\phi}\pi) - \mathcal{H}(x) + J(x)\phi(x) \right] \right\} \\ &= \int [d\phi(x)] \exp \left\{ \int d^4x [\mathcal{L}(x) + J(x)\phi(x)] \right\} = \exp iZ(J),\end{aligned}\quad (52)$$

and  $Z(J)$  is a functional partition function in quantum field theory. The  $n$ -point Green's functions containing only the connected graphs is given by,

$$G_c^{(n)}(x_1, \dots, x_n) = \frac{(-i)^n}{W(J)} \frac{\delta^n W(J)}{\delta J(x_1) \cdots \delta J(x_n)} \bigg|_{J=0} = (-i)^{n-1} \frac{\delta^n Z(J)}{\delta J(x_1) \cdots \delta J(x_n)} \bigg|_{J=0}. \quad (53)$$

Taking the stationary phase approximation of  $W(J)$  naturally results in the classical equations of motions [29, 30]. On the other hand, after integrating out the  $\pi(x)$  field, the functional integrals  $W(I)$  and  $Z(J)$  become covariant. Now, all physical quantities in field theory are, in principle, derivable from  $W(J)$  or  $Z(J)$  in a covariant form, which are standard in text books. I should not repeat these discussion here.

The above formulation is only for bosonic fields. Field theory that describes the real world must also involve fermion (matter) fields. To formulate a similar functional integral for fermionic fields, one needs to introduce fermion coherent states. Similarly, one may try to construct such a coherent state as an eigenstate of the fermion annihilation operator:

$$c_i|\xi\rangle = \xi_i|\xi\rangle, \quad \{c_i, c_j^\dagger\} = \delta_{ij}. \quad (54)$$

However, since the fermion creation and annihilation operators satisfy the anticommutation relationship, the eigenvalue  $\xi_i$  of the annihilation operator is its classical analogy which cannot be an

ordinary number. The quantum-classical corresponding principle simply requires that  $\xi_i$  must be anticommute:

$$\xi_i \xi_j = -\xi_j \xi_i, \quad \xi_i \xi_j^* = \xi_j^* \xi_i, \quad \xi_i^2 = 0. \quad (55)$$

The numbers satisfies the above relations are called Grassmann numbers. Functions of Grassmann numbers are given by

$$f(\xi_i) = f^0 + f_i^{(1)} \xi_i + f_{ij}^{(2)} \xi_i \xi_j + \cdots, \quad (56)$$

and the Grassmann integrals are defined as

$$\int d\xi = 0, \quad \int d\xi \xi = 1. \quad (57)$$

Using these properties, the fermionic coherent state can be written explicitly as

$$|\xi\rangle = \exp \left\{ -\frac{1}{2} \sum_i \xi_i^* \xi_i \right\} \exp \left\{ \xi_i c_i^\dagger \right\} |0\rangle, \quad (58)$$

where  $|0\rangle$  is the fermion vacuum state:  $c_i|0\rangle = 0$ . For the fermionic coherent states, the resolution of identity can be written similarly as

$$\int \prod_i d\xi_i^* \xi_i |\xi\rangle \langle \xi| = I. \quad (59)$$

Based on these properties of fermionic coherent states, the functional integral of fermion fields is rather easy to derive [31]. Consider a fermion field  $\psi$  coupling with a scalar boson field  $\phi$ , the Lagrangian is

$$\mathcal{L}(\bar{\psi}, \psi, \phi) = \bar{\psi}(i \not{\partial} - m)\psi + \frac{1}{2}[(\partial\phi)^2 - m^2\phi^2] - g\phi\bar{\psi}\psi. \quad (60)$$

Following the same procedure, one obtains the functional integral

$$W(\bar{\zeta}, \zeta, J) = \int [d\bar{\xi}][d\xi][d\phi] \exp \left\{ \int d^4x \left[ \mathcal{L}(\bar{\xi}, \xi, \phi) + \bar{\zeta}\xi + \bar{\xi}\zeta + J\phi \right] \right\}, \quad (61)$$

where the fermionic sources  $\zeta^*, \zeta$  are also Grassmann numbers. In fact, Schwinger first introduced such a generating functional for fermion fields, in order to derive the fermion field Green's functions [5].

These results can be extended to quantum electrodynamics and quantum chromodynamics, although the later will be more complicated because of non-abelian gauge fields [32, 33]. In most of text books, the discussions on coherent states in field theory are restricted usually in contents of the above formulation. One may derive from such a formulation almost everything about perturbative field theory, such as Feynman rules, the perturbation expansion, and renormalization analysis, etc. At this point, the functional integral of quantum field theory in terms of coherent states is actually nothing special. It is the standard formulation that one can also obtain from other methods. If the field theory can be treated perturbatively, one can always solve the theory in one or other ways, based on the developments of field theory in the last fifty years. The challenge in field theory we faced today (or in the past three decades since the theory of the strong interaction, namely quantum chromodynamics, was proposed) is in the nonperturbation section. There is no a systematic approach in field theory that one can used to completely solve a nonperturbation problem, such as the vacuum structure in non-abelian gauge theory, or bound state problem in strongly coupling systems. In the next few sections, I will try to illustrate some specific problems to see if the generalized coherent states developed later can play some useful roles to the nonperturbation field theory of strongly interacted systems.

## 5 Squeezed Coherent States and Quantum Fluctuations

The first example I go to discuss is how one may use squeezed states to study the low energy quantum fluctuations in strong interaction theory. Different from what is discussed in the previous section where the content may be found in text books, here I should point out that the formulation presented in this section has actually not been completed yet, and more work remains for further investigations.

The squeezed states are a generalization of photon coherent states. Again, the squeezed states first attracted the attention in quantum optics [35, 36]. In the early development, the principal potential applications of squeezed states are in the field of optical communications and “quantum nondemolition experiments” designed for the detection of gravity waves [37]. Later on, because of the “capacity” of treating quantum fluctuations, squeezed states have been used in various subjects, such as quantum measurement theory, quantum nonlinear dynamics, molecular dynamics, dissipative quantum mechanics as well as in quantum gravity and condensed matter physics.

In quantum optics, the uncertainty principle places a damper on the enthusiasm with which quantum engineers approach the problem of coding and transmitting information by optical means. Specifically, the quantum noise inherent in light beams places a limit on the information capacity of an optical beam. Since the uncertainty principle is a statement about areas in phase space, noise levels in different quadratures are statements about intersections of uncertainty ellipses with these axes. Any procedure which can deform or squeeze the uncertainty circle to an ellipse can in principle be used for noise reduction in one of the quadratures. Such squeezing does not violate the uncertainty principle; rather, it places the larger uncertainty in a quadrature not involved in the information transmission process. A typical procedure for squeezing the error ellipse involves applying a classical source to drive two photon emission and absorption processes in much the same way that single photon processes can be used to generate a coherent state of the electromagnetic field.

For simplification, one may consider a basic Hamiltonian describing two photon processes in a single mode

$$H = \omega \left( a^\dagger a + \frac{1}{2} \right) + f(t) a^{\dagger 2} + f^*(t) a^2. \quad (62)$$

Then, the squeezed state can be obtained by directly solving the time-dependent Schrödinger equation. If the initial state is the photon vacuum, a general solution is

$$|\beta\rangle = \exp \left( \frac{1}{2} \beta a^{\dagger 2} - \frac{1}{2} \beta^* a^2 \right) |0\rangle e^{i\varphi}. \quad (63)$$

This is the squeezed states generated by Eq.(62). If one defines the photon's position and momentum coordinates  $(x, p)$  in terms of the creation and annihilation operators as in the previous section, one can then find that

$$\begin{aligned} \Delta x^2 &= \langle \beta | x^2 | \beta \rangle = \frac{1}{2} \left| \cosh|\beta| + \frac{\beta}{|\beta|} \sinh|\beta| \right|^2, \\ \Delta p^2 &= \langle \beta | p^2 | \beta \rangle = \frac{1}{2} \left| \cosh|\beta| - \frac{\beta}{|\beta|} \sinh|\beta| \right|^2, \end{aligned} \quad (64)$$

and  $\Delta x \neq \Delta p$  but  $\Delta x \Delta p \geq \frac{1}{2}$  (here we set  $\hbar = 1$ ). While the vacuum has a circle uncertainty ( $\Delta x = \Delta p$ ) in phase space. This shows that the operator  $D_{sq}(\beta) = \exp \left( \frac{1}{2} \beta a^{\dagger 2} - \frac{1}{2} \beta^* a^2 \right)$  squeezes the uncertainty circle of a wave packet into an ellipses so that quantum fluctuation (noise) can be reduced in one of the quadratures.

In general, it is desirable to squeeze a field coherent state which can be generated by

$$H = \omega \left( a^\dagger a + \frac{1}{2} \right) + f_2(t) a^{\dagger 2} + f_2^*(t) a^2 + f_1(t) a^\dagger + f_1^*(t) a. \quad (65)$$

The sequence in which the processes of coherent state formation and squeezing occur is governed by the time dependence of the functions  $f_2(t)$  and  $f_1(t)$ . The general form of the state at the time

$t$  can be expressed (apart from a phase factor) by

$$|z\beta\rangle = \exp\left(za^\dagger - z^*a\right) \exp\left(\frac{\beta}{2}a^{\dagger 2} - \frac{\beta^*}{2}a^2\right)|0\rangle, \quad (66)$$

where the complex variables  $z$  and  $\beta$  are functions of the time  $t$  in general. Eq. (66) is usually called squeezed coherent state. The physical process of the squeezed coherent states can be understood as follows: by first squeezing the vacuum (the wave packet) by the two photon excitations, and then displacing it as a photon coherent state by the external source.

Using group theory, one can show that the squeezed coherent states must also form a overcomplete set of states.

$$\int \frac{dzdz^*}{2\pi} \frac{dfdg}{2\pi} |z\beta\rangle \langle z\beta| = I \quad (67)$$

where the variables  $f$  and  $g$  are introduced by Jackiw and Kerman [38] to characterize quantum fluctuations (noise) of the position and momentum:

$$\Delta x^2 = f, \quad \Delta p^2 = \frac{1}{4f} + 4fg^2, \quad (68)$$

These two variables relate to the squeezing parameter  $\beta$  by (64). By the completeness, one can also derive a path integral of quantum mechanics in the squeezed coherent state representation:

$$\begin{aligned} \langle z'(t_f)\beta'(t_f)|z(t_0)\beta(t_0)\rangle &= \int [dx(t)] \left[ \frac{dp(t)}{2\pi} \right] [df(t)] \left[ \frac{dg(t)}{2\pi} \right] \\ &\times \exp \left\{ i \int_{t_0}^{t_f} dt \left[ \frac{1}{2}(p\dot{x} - x\dot{p}) - f\dot{g} - \mathcal{H}_{\text{eff}} \right] \right\}, \end{aligned} \quad (69)$$

where the effective Hamiltonian is the matrix element of the Hamiltonian operator  $[H = p^2/2 + V(x)]$  in the squeezed coherent state:

$$\mathcal{H}_{\text{eff}}(x, p, f, g) = \frac{1}{2}p^2 + \frac{1}{8f} + 2fg^2 + \exp\left(\frac{f}{2}\frac{\partial^2}{\partial x^2}\right)V(x). \quad (70)$$

The expression of path integral in squeezed coherent states shows that  $f$  and  $g$  which characterize quantum fluctuations become a pair of conjugate variables. The extremal values of the exponent in the path integral leads to the following generalized equations of motion:

$$\begin{aligned} \frac{dx}{dt} &= \partial \mathcal{H}_{\text{eff}} / \partial p, & \frac{dp}{dt} &= -\partial \mathcal{H}_{\text{eff}} / \partial x, \\ \frac{df}{dt} &= \partial \mathcal{H}_{\text{eff}} / \partial g, & \frac{dg}{dt} &= -\partial \mathcal{H}_{\text{eff}} / \partial f. \end{aligned} \quad (71)$$

Physically, the equations of motion for  $(x, p)$  determine the time evolution of the center of wave packets and those for  $f, g$  characterize the time evolution of the quantum fluctuations (quadratures). Therefore, the variables  $(f, g)$  describe the squeezing and spreading of quadratures in times, which provides a classical-like dynamical theory for the controlling of quantum noise and signal.

It is worth pointing out that although the concept of squeezed state first attracted the attention in quantum optics, squeezed state itself was introduced much earlier by Valatin and Bulter in the study of superfluidity [4, 39]. Similar to the BCS state of superconductivity (which I will discuss later), Valatin's superfluid ground state is defined as

$$\begin{aligned} |\{z_k\beta_k\}\rangle &= \exp \left\{ \sum_k (z_k a_k^\dagger - z_k^* a_k) \right\} \exp \left\{ \sum_k \frac{1}{2} (\beta_k a_k^\dagger a_{-k}^\dagger - \beta_k^* a_k a_{-k}) \right\} |0\rangle \\ &= \exp \left\{ \sum_k (z_k a_k^\dagger - z_k^* a_k) \right\} |\{\beta_k\}\rangle \end{aligned} \quad (72)$$

which is the standard form of squeezed coherent states one currently used. In many-body picture, such states have two consequences. The squeezed operator  $\exp \left\{ \sum_k \frac{1}{2} (\beta_k a_k^\dagger a_{-k}^\dagger - \beta_k^* a_k a_{-k}) \right\}$  acting on the trivial vacuum generates a canonical transformation of quasiparticles:

$$\alpha_k = \cosh |\beta_k| a_k - \frac{\beta_k}{|\beta_k|} \sinh |\beta_k| a_{-k}^\dagger, \quad \alpha_k |\{\beta_k\}\rangle = 0. \quad (73)$$

Then using the quasiparticle vacuum to generate a bosonic coherent state. With such a state, one may develop a microscopic theory of superfluid helium, in which the normal and superfluid states become a direct analogy of noise and signal in partially coherent radiation fields.

We now use the squeezed coherent state to formulate a possible theory that may be useful in addressing the low energy quantum fluctuations in field theory. Let us consider again the neutral scalar field ( $\phi^4$  theory) as an example. One can define the squeezed coherent state of the field  $\phi$  as [40]

$$\begin{aligned} |\Psi\rangle &= \mathcal{N} \exp \left\{ i \int d^3x [\pi(x) \phi_{op}(x) - \phi(x) \pi_{op}(x)] \right\} \\ &\times \exp \left\{ \int d^3x d^3y [\phi_{op}(x) D(x, y) \phi_{op}(y)] \right\} |0\rangle \end{aligned} \quad (74)$$

where  $\mathcal{N}$  is a normalization constant. This squeezed coherent state is also defined at a given instant time so that  $t = t_x = t_y$ . One can show that:

$$\begin{aligned} \langle \Psi | \phi_{op}(x) | \Psi \rangle &= \phi(x), \quad \langle \Psi | \pi_{op}(x) | \Psi \rangle = \pi(x), \\ \langle \Psi | \phi_{op}(x) \phi_{op}(y) | \Psi \rangle &= \phi(x) \phi(y) + \Phi(x, y), \\ \langle \Psi | \pi_{op}(x) \pi_{op}(y) | \Psi \rangle &= \pi(x) \pi(y) + \frac{1}{4} \Phi^{-1}(x, y) \\ &+ 4 \int d^3x' d^3y' \Pi(x, x') \Phi(x', y') \Pi(y', y), \end{aligned} \quad (75)$$

where

$$\begin{aligned} D(x, y) &\equiv \frac{1}{2} [\Phi_0^{-1}(x, y) - \Phi^{-1}(x, y)] + 2i\Pi(x, y), \\ \Phi_0(x, y) &= \langle 0 | \phi_{op}(x) \phi_{op}(y) | 0 \rangle. \end{aligned} \quad (76)$$

The squeezed coherent states of bosonic field are also overcomplete, namely

$$\int [d\phi(x)] \left[ \frac{d\pi(x)}{2\pi} \right] [d\Phi(x, y)] \left[ \frac{d\Pi(x, y)}{2\pi} \right] |\Psi\rangle \langle \Psi| = I. \quad (77)$$

Following the similar procedure discussed in the previous section, one can derive the functional integral  $W(J)$  in the squeezed coherent state representation:

$$\begin{aligned} W(J) &= \int [d\phi(x)] \left[ \frac{d\pi(x)}{2\pi} \right] [d\Phi(x, y)] \left[ \frac{d\Pi(x, y)}{2\pi} \right] \\ &\times \exp \left\{ \int d^4x \left[ \frac{1}{2} (\pi \dot{\phi} - \dot{\phi} \pi) - \Phi \dot{\Pi} - \mathcal{H}_{\text{eff}}(x) + J(x) \phi(x) \right] \right\}, \end{aligned} \quad (78)$$

where

$$\begin{aligned} \mathcal{H}_{\text{eff}} &= \frac{1}{2} [\pi^2(x) + (\nabla \phi)^2 + m^2 \phi^2] + \exp \left( \frac{\Delta(x)}{2} \frac{\partial^2}{\partial \phi^2} \right) V(\phi) \\ &+ \frac{1}{8} \Phi^{-1}(x, x) + 2 \int d^3x' d^3y' \Pi(x, x') \Phi(x', y') \Pi(y', x) \\ &+ \frac{1}{2} \left[ \lim_{x \rightarrow y} (\nabla_x \nabla_y \Phi(x, y)) + m^2 \Phi(x, x) \right] \end{aligned} \quad (79)$$



and  $\Delta(x, t) = \lim_{x \rightarrow y} [\Phi(x, y) - \Phi_0(x, y)]$ . Here I have not integrated out the conjugate momentum  $\{\pi(x), \Pi(x)\}$  to obtain a covariant functional integral. The physical picture of this new formulation is that besides the original field variable  $\phi(x), \pi(x)$  in the Hamiltonian formulation, quantum fluctuations, characterized by  $\Phi(x, y)$  and  $\Pi(x, y)$ , are introduced as new dynamical field variables. These new dynamical field variables describe the low energy excitations (i.e. the **composite particles**) of strongly interacted systems. Similarly, taking the extreme value of the exponent in (78), one can find the equation of motion that determine the classical-like solution  $\phi_0$ :

$$(\partial_\mu^2 - m^2)\phi_0 + \exp\left(\frac{\Delta(x)}{2} \frac{\partial^2}{\partial \phi_0^2}\right) V'(\phi_0) = 0, \quad (80)$$

which is coupled with the composite field  $\Phi$ . The equation of motion for its quantum fluctuations by  $\Phi$  is much more complicated.

Usefulness of the squeezed state functional integral is that one can derive an effective theory for the low energy composite particle fields coupling with the original fields. Here I may propose a procedure how to develop such an effective theory. First, one can determine the “classical” ground state by minimizing the effective Hamiltonian with respect to variables  $\phi, \pi$  as well as  $\Phi, \Pi$ , which results in  $\phi_0, (\pi_0 = 0)$  and  $\Phi_0, (\Pi_0 = 0)$ . Then, expanding the effective Lagrangian near  $(\phi_0, \pi_0, \Phi_0, \Pi_0)$  up to the second-order, namely only keeping the quadratic terms in  $(\delta\phi, \delta\pi, \delta\Phi, \delta\Pi)$ . Quantum effects become the time-dependent fluctuations about the classical ground states. The corresponding linearized equations of motion determine the dispersions of quasiparticles and composite particles, denoted by  $\omega_k$  and  $\gamma_k$ , respectively. For strong interaction field, usually  $\omega_k > \gamma_k$  (due to the spontaneously symmetry breaking). Thus, in the low energy scale ( $\omega_k > \mu \geq \gamma_k$ ), the composite particles and the quasiparticles are decoupled. Only the composite particles are kept with the high order corrections (as a perturbation) to form the low energy effective Lagrangian. In the intermediate energy scale ( $\mu \sim \omega_k$ ), both the composite particles and quasiparticles become active and coupled each other. The effective theory is then determined by the Lagrangian of the composite particles coupled with quasiparticle degrees of freedom. In a rather high energy scale ( $\mu \gg \omega_k$ ),  $\Phi$  and  $\Pi$  should spread averagely over the entire space-time space such that only the original field variables remain. Thus, the theory returns back to the original one in high energy region.

The above procedure is different from the conventional procedure of constructing a low energy effective theory. In the conventional approach, the low energy effective theory is constructed by separating the field variables into the low energy and high energy parts. Then, using the functional integral discussed in the previous section to integrate out the high energy part. The resulting Lagrangian is an effective Lagrangian for the low energy physics. The advantage of the conventional approach is that one can use the powerful renormalization group analysis of Wilson [41] to extract universal scaling properties contained in the theory, without explicitly solving the theory itself. However, in reality, physical degrees of freedoms must also be very different in different energy scales. A typical example is the strong interaction in which the degrees of freedom are quarks and gluons in high energy region. But in low energy region, the degrees of freedom become hadrons which are composite particles of quarks and gluons. Thus, in the conventional approach, the effective theory cannot catch the right physical degrees of freedom. Therefore, beyond the critical phenomena, the conventional approach may have its limitation in applications. The squeezed coherent state formulation of the functional integral may provide a new method for the developments of effective field theory, although at this point a lot of work remains for further investigations.

A potential application of the squeezed coherent states in field theory is the Yang-Mills gauge theory, especially the color SU(3) gauge theory in quantum chromodynamics. Of course, the situation in quantum chromodynamics is much more complicated. Because of the nonlinear properties in non-abelian gauge theory, the conventional functional integral is already quite complicated. However, the conventional functional integral in non-abelian gauge theory is only useful for the derivation of covariant Feynman rules and the analysis of renormalizability. In other words, it is only useful for perturbation calculations. As it is well-known, the difficulty of QCD lies in its non-perturbation domain, where quantum fluctuation must be strong. Furthermore, the field strength

of non-abelian gauge contains the single as well as double gauge boson emissions and absorptions:

$$F_{\mu\nu}^a = \partial_\mu A_\nu^a - \partial_\nu A_\mu^a - gf^{abc}A_\mu^b A_\nu^c. \quad (81)$$

The squeezed coherent states may be the natural quantum states describing non-abelian gauge fields. If one can complete and extend the above formulation to the non-abelian gauge theory, then it may be able to develop a low energy effective theory for non-abelian gauge fields. I believe that such a low energy effective theory should be capable in dealing with gluon condensation, gluball states as well as low energy gluon dynamics.

## 6 Spin Coherent States and Non-Linear Sigma Model

So far, I have only discussed bosonic-type coherent states in field theory. However, coherent states, in terms of the language of group theory, are embedded in a topologically nontrivial geometrical space which involve a deep implication in physics [13]. The simplest coherent state carried a topologically nontrivial geometrical space is the *spin coherent state*. The most attractive property in spin coherent states is that its topological structure naturally induces Dirac's **magnetic monopole** [34]. Meanwhile, the spin coherent state representation of the path integral for a multi-spin system gives a realistic realization of Non-Linear Sigma Model which is an important field theory model in condensed matter physics and particle physics. Thus, before I go to discuss the general physics implication containing in the geometrical structure of coherent states, it may be useful to illustrate first the spin coherent state in details.

Let us start with a simple example: a spin-1/2 particle in a varying magnetic field:  $B(t) = (B_x(t), B_y(t), B_z(t))$ , described by the Hamiltonian,

$$H(t) = -\mu \vec{S} \cdot B(t). \quad (82)$$

Here  $\mu$  is the particle's magnetic moment,  $\vec{S} = (S_x, S_y, S_z)$  the spin operator that satisfies the usual angular momentum commutation relationship. The evolution of system governed by (82) can be determined by the Schrödinger equation, whose general solution can be written as

$$|\psi(t)\rangle = \alpha(t)|\downarrow\rangle + \beta(t)|\uparrow\rangle. \quad (83)$$

Substituting (83) into Schrödinger equation, it is easy to determine the time-dependence of the parameters  $\alpha(t)$  and  $\beta(t)$ . However, in order to derive the spin coherent state, here, I may only concentrate on the structure of the state (83). The normalization of (83) results in a constraint on the parameters  $\alpha(t)$  and  $\beta(t)$ :

$$|\alpha(t)|^2 + |\beta(t)|^2 = 1. \quad (84)$$

If I parameterize  $\beta = \sin \frac{\theta}{2} e^{-i\varphi}$ , Eq. (83) can be expressed as

$$|\psi(t)\rangle = \left( \cos \frac{\theta(t)}{2} + \sin \frac{\theta(t)}{2} e^{-i\varphi(t)} S^+ \right) |\downarrow\rangle e^{i\phi(t)}, \quad (85)$$

the raising and lowering spin operators  $S^\pm$  are defined by  $S^\pm = S_x \pm iS_y$ , and  $S^+|\downarrow\rangle = |\uparrow\rangle$ ,  $S^-|\uparrow\rangle = |\downarrow\rangle$ . Furthermore, one can easily show that

$$\left( \cos \frac{\theta}{2} + \sin \frac{\theta}{2} e^{-i\varphi} S^+ \right) |\downarrow\rangle = \exp \left\{ \frac{\theta}{2} e^{-i\varphi} S^+ - \frac{\theta}{2} e^{i\varphi} S^- \right\} |\downarrow\rangle \equiv |\theta\varphi\rangle. \quad (86)$$

Then, (85) can be simply expressed as

$$|\psi(t)\rangle = |\theta(t)\varphi(t)\rangle e^{i\phi(t)}. \quad (87)$$

The state  $|\theta\varphi\rangle$  is a standard expression of the spin coherent states.

As I mentioned a very important property of the spin coherent state is that  $|\theta\varphi\rangle$  is embedded in a topologically nontrivial geometrical space, i.e., a two-dimensional sphere  $S^2$ . This can be varified directly from (85). Since  $S_i = \sigma_i/2$  where  $\sigma_i$  is Pauli matrix,

$$\begin{aligned} D_s(\eta) &\equiv \exp\{\eta S^+ - \eta^* S^-\} = \exp \begin{bmatrix} 0 & \eta \\ -\eta^* & 0 \end{bmatrix} \\ &= \begin{bmatrix} \cos|\eta| & \frac{\eta}{|\eta|} \sin|\eta| \\ -\frac{\eta^*}{|\eta|} \sin|\eta| & \cos|\eta| \end{bmatrix} = \begin{bmatrix} x_0 & x \\ -x^* & x_0 \end{bmatrix} \end{aligned} \quad (88)$$

and  $x_0$  is real while  $x = x_1 + ix_2$ . Also,  $D_s(\eta)$  is a unitary operator which leads to

$$x_0^2 + |x|^2 = x_0^2 + x_1^2 + x_2^2 = 1. \quad (89)$$

In other words, the parameter space of  $D_s(\eta)$  is a two-sphere  $S^2$  with  $\eta = \frac{\theta}{2}e^{-i\varphi}$ . Therefore, the spin coherent states are ono-to-one corresponding to the points on  $S^2$  except for the north pole where it is ambiguous since all values of  $\varphi$  correspond to the same point.

However, in defining the above spin coherent states, there is an ambiguous. For example, one can also define the spin coherent states as

$$|\theta\varphi\rangle' = \left( \cos\frac{\theta}{2}e^{i\varphi} + \sin\frac{\theta}{2}S^+ \right) |\downarrow\rangle, \quad (90)$$

which are also ono-to-one corresponding to the points in  $S^2$  but except for the south pole. These two spin coherent states are related simply by a phase factor,

$$|\theta\varphi\rangle' = e^{i\varphi} |\theta\varphi\rangle. \quad (91)$$

Geometrically, these two coherent states define the two “patches” of  $S^2$ . Since these two states are only different by a phase factor, quantum mechanically, they must be equivalent. This implies that there is a gauge degree of freedom in the spin coherent states.

To see clearly the physical implication induced by the topological structure of spin coherent states, one can construct the path integral of a quantum spin system. The spin coherent state also obeys the overcompleteness:

$$\int d\mu(\theta\varphi) |\theta\varphi\rangle \langle\theta\varphi| = I, \quad (92)$$

where  $d\mu(\theta\varphi) = \sin\theta d\theta d\varphi/2\pi$  is an invariant measure on  $S^2$ . Then, it is easy to show that the path integral of quantum mechanics for  $H = H(\vec{S})$  is given by

$$\langle\theta'(t_f)\varphi'(t_f)|\theta(t_0)\varphi(t_0)\rangle = \int [d\mu(\theta\varphi)] \exp \left\{ i \int_{t_0}^{t_f} dt \left[ \langle\theta\varphi| i \frac{d}{dt} |\theta\varphi\rangle - \langle\theta\varphi| H(\vec{S}) |\theta\varphi\rangle \right] \right\}. \quad (93)$$

In this path integral, the first term in the exponent,

$$\omega[\theta\varphi] \equiv \int_{t_0}^{t_f} dt \langle\theta\varphi| i \frac{d}{dt} |\theta\varphi\rangle = \int_{\varphi_0}^{\varphi_f} \frac{1}{2} (1 - \cos\theta) d\varphi, \quad (94)$$

is pure geometric that only depends on the trajectory over the sphere, but not on its explicit time dependence. For a closed path,  $\omega[\theta\varphi]$  is actually a **Berry phase** of the spin history [43]. Therefore,  $\omega[\theta\varphi]$  is a gauge invariant one-form defined on the sphere  $S^2$ :

$$\omega[\theta\varphi] = \int_{\varphi_0}^{\varphi_f} A_\varphi d\varphi = \int_{\varphi_0}^{\varphi_f} \mathcal{A}(\mathbf{n}) d\mathbf{n}, \quad (95)$$

where  $\mathbf{n} = (\sin\theta \cos\varphi, \sin\theta \sin\varphi, \cos\theta)$  is a unit vector, and  $\mathcal{A}(\mathbf{n})$  is a unit vector potential. Compare with (94) and (95), one can find that

$$\mathcal{A}^a = \frac{1 - \cos\theta}{2 \sin\theta} \hat{\varphi}. \quad (96)$$

This vector potential has one singularity at the south pole. It is this singularity where the Dirac string which carries the magnetic monopole flux enters the sphere. Hence,  $\mathcal{A}^a$  is nothing but a gauge potential of Dirac's magnetic monopole. Similarly, for the spin coherent state  $|\theta\varphi\rangle'$ , the corresponding gauge potential is

$$\mathcal{A}^b = \frac{1 + \cos\theta}{2 \sin\theta} \hat{\varphi}. \quad (97)$$

$\mathcal{A}^a$  and  $\mathcal{A}^b$  define the two non-singular patches of the monopole section. Their difference is a pure  $U(1)$  gauge in the overlapping equatorial region,  $S^1$ ,

$$\mathcal{A}_\varphi^b = \mathcal{A}_\varphi^a + d\varphi = \mathcal{A}_\varphi^a - ig^{-1}dg. \quad (98)$$

where  $g = e^{i\varphi} \in U(1)$ .

The existence of the above gauge degrees of freedom can be understood clearly by looking at the general definition of coherent states based on group theory [9, 10]. In group theory, quantum states of a spin system form a unitary representation  $V^s$  of the  $SU(2)$  group, here  $s$  is an arbitrary spin. Choosing a fixed state, such as the lowest-weight state  $|ss_z\rangle = |s-s\rangle \in V^s$ , one can define spin coherent states as

$$|g\rangle_s = g|s-s\rangle, \quad g \in SU(2). \quad (99)$$

In general,  $g = \exp(i\alpha S_x) \exp(i\beta S_y) \exp(i\gamma S_z) = \exp(\frac{\theta}{2}e^{-i\varphi}S^+ - \frac{\theta}{2}e^{i\varphi}S^-) \exp(i\gamma' S_z)$ . Note that this decomposition, called the Baker-Campbell-Hausdorff formula, is unique. As a result, one can rewrite the above spin coherent states as

$$|g\rangle_s = \exp(\frac{\theta}{2}e^{-i\varphi}S^+ - \frac{\theta}{2}e^{i\varphi}S^-) \exp(i\gamma' S_z)|s, -s\rangle = |\theta\varphi\rangle e^{i\chi}. \quad (100)$$

where

$$|\theta\varphi\rangle = \exp(\frac{\theta}{2}e^{-i\varphi}S^+ - \frac{\theta}{2}e^{i\varphi}S^-)|s, -s\rangle \quad (101)$$

is the standard definition of spin coherent state for an arbitrary spin  $s$  [42]. Spin  $s = 1/2$  discussed above is a special case. As one can see, apart from a phase factor, the spin coherent states can be generated by a unitary spin rotational operator acting on the fixed state  $|s-s\rangle$ . The unitary operator  $\exp(\frac{\theta}{2}e^{-i\varphi}S^+ - \frac{\theta}{2}e^{i\varphi}S^-)$  is the coset representation of the space  $SU(2)/U(1) \simeq S^2$ . Therefore the sphere  $S^2$  determines the topological structure of spin coherent states. The magnetic monopole potential  $\mathcal{A}^a$  defines a  $U(1)$  fibre bundle over this sphere  $S^2$ . Meanwhile, the spin coherent states contain an arbitrary phase  $\chi$ . In quantum mechanics, a quantum state is specified up to a phase factor, namely, physics is invariant for different choices of such a phase factor so that this phase factor is usually ignored. However, when quantum states are embedded in a topologically nontrivial space, this phase freedom is indeed the associated gauge degrees of freedom of the fibre bundle over the space. In spin coherent states, the phase  $\chi$  is just the gauge degree of freedom that connects different choices of magnetic monopole potentials. Ignoring (or fixing) this phase factor corresponds to a gauge fixing.

Furthermore, the topological properties of the spin coherent states also play an important role in the study of spin dynamics. A typical example is the Heisenberg model in condensed matter physics. Heisenberg model is used to understand quantum magnetism of strongly correlated electron systems. The model Hamiltonian considered here is very simple:

$$H_J = J \sum_{\langle i,j \rangle} \mathbf{S}_i \cdot \mathbf{S}_j \quad (102)$$

which describes a many-spin (each spin =  $s$ ) system with the nearest neighbor exchange interaction. "Classically", the ground state of the above Hamiltonian is easily determined. When  $J > 0$ , the minimum energy is given by the state in which the nearest-neighbor spins are always anti-alignment. These states are called in literatures the Néel states. Correspondingly, the system is an antiferromagnet. If  $J < 0$ , the ground state is simply given by the state with all spin aligned in

the same direction, which is a ferromagnetic state. These consequences can be obtained explicitly by taking the spin coherent state

$$|\{\theta_i \varphi_i\}\rangle = \prod_i \exp \left\{ \frac{\theta_i}{2} e^{-i\varphi_i} S_i^+ - \frac{\theta_i}{2} e^{i\varphi_i} S_i^- \right\} |s - s\rangle \quad (103)$$

as a trial wave function and minimizing the model Hamiltonian

$$\delta(\langle \{\theta_i \varphi_i\} | H_J | \{\theta_i \varphi_i\} \rangle) = \delta(Js^2 \sum_{\langle i,j \rangle} [\cos \theta_i \cos \theta_j + \sin \theta_i \sin \theta_j \cos(\varphi_i - \varphi_j)]) = 0. \quad (104)$$

The resulting ground state is given by

$$\begin{cases} J > 0 \rightarrow \theta_{i+1} = \pi - \theta_i, \varphi_{i+1} = \varphi_i + \pi \rightarrow \text{antiferromagnet} \\ J < 0 \rightarrow \theta_{i+1} = \theta_i, \varphi_{i+1} = \varphi_i \rightarrow \text{ferromagnet} \end{cases} \quad (105)$$

An important concept one can obtain from the above result is that the ground state spontaneously breaks the global spin rotational symmetry. As one can check the model Hamiltonian is invariant under global spin rotational transformations:  $T = \exp(i\vec{\alpha} \cdot \mathbf{S})$ , where  $\mathbf{S} = \sum_i \mathbf{S}_i$ . While the ground state energy does not change when all the spins in (105) are globally rotated. This leads to a SO(3) degeneracy of the ground states, namely these ground states have a lower symmetry than the Hamiltonian. Such a situation is called the **spontaneously symmetry breaking**. Quantum mechanically, it leads to gapless spin-wave excitations in the Heisenberg model, and Goldstone bosons in general.

The quantum dynamics of interacting spins can be studied from the time-evolution of the system at zero-temperature. The time-evolution is determined by the Green's function which is defined by the matrix element of the evolution operator between two spin coherent states:

$$G(t_f, t_0) = \int \prod_i [d\mu(\theta_i \varphi_i)] \exp \{ iS[\theta_i(t), \varphi_i(t)] \}, \quad (106)$$

here, the effective action is given by

$$\begin{aligned} S[\theta_i(t), \varphi_i(t)] &= \int_{t_0}^{t_f} dt \left[ \langle \{\theta_i \varphi_i\} | i \frac{d}{dt} | \{\theta_i \varphi_i\} \rangle - \langle \{\theta_i \varphi_i\} | H_J | \{\theta_i \varphi_i\} \rangle \right] \\ &= \int_{t_0}^{t_f} dt \left[ s \sum_i (1 - \cos \theta_i) \frac{d\varphi_i}{dt} \right. \\ &\quad \left. - Js^2 \sum_{\langle i,j \rangle} [\cos \theta_i \cos \theta_j + \sin \theta_i \sin \theta_j \cos(\varphi_i - \varphi_j)] \right] \end{aligned} \quad (107)$$

Note that the thermal dynamics can be obtained in a similar form. The partition function can be expressed in terms of a spin coherent state path integral as well:

$$\begin{aligned} \mathcal{Z}(\beta) &= \langle \{\theta_i \varphi_i\} | \exp \{ -\beta H_J \} | \{\theta_i \varphi_i\} \rangle \\ &= \int [d\mu(\beta_i(\tau) \varphi_i(\tau))] \exp \{ -S[\theta_i(\tau), \varphi_i(\tau)] \} \end{aligned} \quad (108)$$

with

$$\begin{aligned} S[\theta_i(\tau), \varphi_i(\tau)] &= \int_0^\beta d\tau \left[ -is \sum_i (1 - \cos \theta_i) \frac{d\varphi_i}{d\tau} \right. \\ &\quad \left. + Js^2 \sum_{\langle i,j \rangle} [\cos \theta_i \cos \theta_j + \sin \theta_i \sin \theta_j \cos(\varphi_i - \varphi_j)] \right], \end{aligned} \quad (109)$$

where  $\tau$  is an imaginary time from 0 to  $\beta = 1/kT$ .

Eq. (107) shows that spin dynamics is induced by the geometrical phase  $\omega = \int dt \sum_i s(1 - \cos \theta_i) \dot{\varphi}_i$  which contains the time derivative and therefore leads to the equation of motion for  $\{\theta_i(t), \varphi_i(t)\}$ . It also shows that the magnetic monopole potential  $A_{\varphi_i}$  is actually the conjugate momentum of  $\varphi_i$  in spin dynamics. If one defines the generalized position and momentum coordinates by

$$q_i = \varphi_i, \quad p_i = s(1 - \cos \theta_i), \quad (110)$$

and expands the effective action  $\mathcal{S}[\theta_i(t), \varphi_i(t)]$  around the ground state  $\{q_i^0, p_i^0, q_{i+1}^0, p_{i+1}^0\}$  [given by Eq. (105)]:

$$\begin{aligned} q_i &= q_i^0 + \delta q_i, & p_i &= p_i^0 + \delta p_i, \\ q_{i+1} &= q_{i+1}^0 + \delta q_{i+1}, & p_{i+1} &= p_{i+1}^0 + \delta p_{i+1}, \end{aligned} \quad (111)$$

up to the quadratic terms, one can determine explicitly the dispersion of **spin-wave excitations** [44, 45].

However, since  $p_i$  is related to the magnetic monopole potential, the conjugate coordinates  $\{q_i, p_i\}$  used above are indeed gauge dependent quantities. To explore the possible topological effects in spin dynamics, it is better to use a gauge invariant formulation. This may be done by using a global notation  $\mathbf{n}$  (a unit vector) to represent the spin direction, without specifying the parameterization of the sphere by  $\theta, \varphi$ . Then, the Green's function can be expressed as:

$$G(t_f, t_0) = \int [d\mu(\mathbf{n}_i)] \exp \left\{ i \int_{t_0}^{t_f} dt \left[ s \sum_i \mathcal{A}_i \cdot \dot{\mathbf{n}}_i - Js^2 \sum_{\langle i, j \rangle} \mathbf{n}_i \cdot \mathbf{n}_j \right] \right\}. \quad (112)$$

Taking the continuum limit

$$\mathbf{n}_i \rightarrow c_i \mathbf{n}(\mathbf{x}_i), \quad (113)$$

where  $c_i = 1$  ( $e^{i\mathbf{x}_i \cdot \vec{\pi}}$ ) for the ferromagnet (antiferromagnet),  $|\mathbf{n}(\mathbf{x}_i)| = 1$ , and  $\vec{\pi} = (\pi, \dots, \pi)$ ,  $\mathbf{x}_i \in R^d$ . Then the Green's function can be expressed as

$$G(t_f, t_0) = \int [d\mu(\mathbf{n}(\mathbf{x}))] \exp \left\{ i \int d^{d+1} \mathcal{L}(\mathbf{n}) \right\}, \quad (114)$$

where  $\mathcal{L}(\mathbf{n})$  is an effective Lagrangian. In the low energy (long wave-length) limit, it is reduced to the Lagrangian of the **Non-Linear Sigma Model** in  $d+1$ -dimensional space [46],

$$\mathcal{L}(\mathbf{n}) = \frac{1}{2g} \partial_\mu \mathbf{n} \cdot \partial^\mu \mathbf{n} + \dots \quad (115)$$

where “...” denotes the high order derivatives. This Lagrangian ensures the existence of gapless spin-wave excitations. Such a Non-Linear Sigma Model has been widely studied in condensed matter physics, including the problems in quantum magnetism, quantum Hall effect and disorder dynamics.

## 7 Generalized Coherent States and Nonabelian Gauge Fields

In this last section, I will discuss the generalized coherent states and their potential applications in field theory. Generalization of coherent states is based on group theory developed by Perelomov and also Gilmore [9, 10]. The spin coherent state discussed in the previous section is an example of such generalization. Actually, Glauber had pointed out in his seminal paper [7] that the photon coherent states can be constructed starting from any one of three mathematical definitions.

- Definition 1. The coherent states  $|z\rangle$  are eigenstates of the annihilation operator  $a$ :

$$a|z\rangle = z|z\rangle. \quad (116)$$

- Definition 2. The coherent states  $|z\rangle$  are quantum states with a minimum uncertainty relationship:

$$\Delta x^2 \Delta p^2 = \frac{\hbar^2}{4} \quad (117)$$

- Definition 3. The coherent states  $|z\rangle$  can be obtained by applying a displacement operator  $D(z)$  on the ground state of harmonic oscillator:

$$|z\rangle = D(z)|0\rangle, \quad D(z) = \exp(za^\dagger - z^*a). \quad (118)$$

We have analyzed these definitions and pointed out [Zha90] that the generalization of eigenstates of the lowering operator is not always possible. Indeed, the adoption of this definition to generalize the coherent state concept has two major drawbacks: a). Coherent states cannot be defined in Hilbert spaces of finite dimensionality in this way, as we have seen for the spin systems. b). The states so defined do not correspond to physically realizable states, except under special circumstances that the commutator of the annihilation operator (or lowering step operator) and its hermitian adjoint is a multiple of the identity operator. Therefore, under this condition one restricts oneself to the bosonic field. As a result, the generalization based on definition 1 to other dynamical systems is not always applicable.

On the other hand, the generalization based on the definition 2 is by no means unique. The bosonic (photon) coherent states are the minimum uncertainty states essentially because they are non-spreading wave packets. Although the minimum uncertainty states are physically very interesting, the generalization along this direction has several limitations: a). These coherent states can only be constructed for the classically integrable systems in which there exists a set of canonical coordinates and momenta such that the respective Hamiltonians can be reduced to quadrature. This condition requires a flatness condition on the operator algebra which reduce the commutation relations to those of the standard bosonic creation and annihilation operators. b). The wave packets with the minimum uncertainty are not unique. Different ones may have different properties. Also, such states may be incomplete, or even if they are complete it is not certain that the standard form of a resolution of unity exists. Thus, minimum uncertainty states appear to have few, if any, useful properties.

In literatures, the realization of generalized coherent states are indeed achieved based on displacement operators. The basic theme of this development was to intimately connect coherent states with dynamical symmetry groups of a physical problem. Since all physical problems formulated in quantum theory have a dynamical group (although sometimes the group may be too large to be useful), an important outcome of this recognition is that coherent states can be generalized to all the quantum problems.

I should outline here a generalization procedure how an arbitrary coherent state can be generated by displacement operators. Consider a set of operators  $\{T_i\}$  closed under commutation:

$$[T_i, T_j] = T_i T_j - T_j T_i = \sum_k C_{ij}^k T_k, \quad (119)$$

That is,  $\{T_i\}$  span a algebra  $\mathfrak{g}$ , and  $C_{ij}$  in (119) are structure constants of  $\mathfrak{g}$ . If  $\mathfrak{g}$  is a semisimple Lie algebra, it is more convenient to write  $\{T_i\}$  in terms of the standard Cartan basis  $\{H_i, E_\alpha, E_\alpha^\dagger = E_{-\alpha}\}$ :

$$\begin{aligned} [H_i, H_j] &= 0, & [H_i, E_\alpha] &= \alpha_i E_\alpha, \\ [E_\alpha, E_{-\alpha}] &= \alpha^i H_i, & [E_\alpha, E_\beta] &= N_{\alpha\beta} E_{\alpha+\beta}. \end{aligned} \quad (120)$$

In quantum theory, for such a given set of closed operators  $\{T_i\}$ , the quantum states are described by a Hilbert space  $V^\Lambda$  which is a representation of  $\mathfrak{g}$ . Let  $G$  be the covering group of  $\mathfrak{g}$ . The Hilbert space  $V^\Lambda$  carries a unitary irreducible representation  $\Gamma^\Lambda$  of  $G$ . One may choose a normalized state  $|\phi_0\rangle$  in the Hilbert space  $V^\Lambda$  as a fixed state. Then the generalized coherent state is generated by an element  $g \in G$  acting on the fixed state  $|\phi_0\rangle$ .

$$|g\rangle_G = g|\phi_0\rangle. \quad (121)$$

In group theory, every element  $g \in \mathbf{G}$  can be uniquely decomposed into a product of two group elements:  $g = kh$ , here one should require  $h \in \mathbf{H}$  such that

$$h|\phi_0\rangle = |\phi_0\rangle e^{i\chi}, \quad (122)$$

and  $\mathbf{H}$  is the maximum subgroup of  $\mathbf{G}$  that leaves the fixed state invariant up to a phase factor. While  $k$  is an operator of the coset space  $\mathbf{G}/\mathbf{H}$ . If  $\mathbf{G}$  is a semisimple Lie group and  $|\phi_0\rangle$  is the lowest weight state,  $k$  can be generally written as

$$k \equiv D_G(\eta) = \exp \left\{ \sum_{\alpha>0} (\eta_\alpha E_\alpha - \eta^* E_{-\alpha}) \right\} \in \mathbf{G}/\mathbf{H}. \quad (123)$$

This operator  $D_G(\eta)$  is usually called a displacement operator of  $\mathbf{G}$ , which gives a coset representation of  $\mathbf{G}/\mathbf{H}$ . As a result,

$$|g\rangle_{\mathbf{G}} = D_G(\eta)|\phi_0\rangle e^{i\chi} = |\Phi(Z)\rangle e^{i\chi}, \quad (124)$$

Perelomov and Gilmore [9, 10] define the state  $|\Phi(Z)\rangle$  as the generalized coherent states of  $\mathbf{G}$ :

$$|\Phi(Z)\rangle = D_G(\eta)|\phi_0\rangle = \mathcal{N}(Z) \exp \left\{ \sum_{\alpha>0} Z_\alpha E_\alpha \right\} |\phi_0\rangle, \quad (125)$$

and  $\mathcal{N}(Z)$  is a normalized constant. The generalized coherent states defined in such a way have two important properties

- The set of the generalized coherent states satisfies:

$$\int d\mu(Z) |\Phi(Z)\rangle \langle \Phi(Z)| = I, \quad (126)$$

where  $d\mu(Z)$  is the  $\mathbf{G}$ -invariant Haar measure on  $\mathbf{G}/\mathbf{H}$ .

- The generalized coherent states are one-to-one corresponding to the points in the coset space  $\mathbf{G}/\mathbf{H}$  except for some singular points, such as the north pole or south pole of the two-sphere in spin coherent states. Therefore, the generalized coherent states are embedded into a topologically nontrivial geometrical space.

Systems discussed in the previous sections are only some simple examples of the generalized coherent states. The harmonic oscillator admits a dynamical group  $H_4$ , called Heisenberg-Weyl group. The photon coherent states are obtained via a one-to-one correspondence with the geometrical coset space  $H(4)/U(1) \times U(1)$  (a complex plane) by the displacement operator  $D(z) \in H(4)/U(1) \times U(1)$ . The two-photon processes has a  $SU(1,1)$  dynamical group. The squeezed states are obtained by the displacement (squeezed) operator  $D_{sq}(\beta) \in SU(1,1)/U(1)$  (a two-dimensional hyperboloid space). And the spin coherent states discussed in the previous section are generated by the displacement operator  $D_s(\theta\varphi) \in SU(2)/U(1)$  (a two-dimensional sphere).

I should emphasize here that the phase  $\chi$  in the group-theoretical coherent state (124) is the  $\mathbf{H}$ -gauge degrees of freedom over the coset space  $\mathbf{G}/\mathbf{H}$ . All the three sets of coherent states discussed in the previous sections contain an  $U(1)$  gauge, but only the sphere (spin) carries a nontrivial fibre bundle so that the gauge degrees of freedom become important. To obtain a non-abelian gauge, one must consider the generalized coherent states of a group  $\mathbf{G}$  whose rank is larger than one such that  $\mathbf{H}$  can be a non-abelian group.

To examine non-abelian gauge degrees of freedom in the generalized coherent states, one may extend the path integral to the generalized coherent state representation. The Green's function is now defined as the matrix element of the evolution operator in the generalized coherent states:

$$G(t_f, t_0) = \langle \Phi'(Z) | T \exp \left\{ -i \int_{t_0}^{t_f} H(t) dt \right\} | \Phi(Z) \rangle. \quad (127)$$



Following the same procedure as it has been done in the previous sections that divides the time interval  $t_f - t_0$  into  $N$  intervals, each with  $\varepsilon = (t_f - t_0)/N$ , then inserts the resolution of identity (126) at each interval point, and finally lets  $N$  go to infinity, the Green's function can be expressed as a generalized coherent state path integral.

$$\begin{aligned} G(t_f, t_0) &= \lim_{N \rightarrow \infty} \int \left( \prod_{i=1}^{N-1} d\mu_i(Z) \right) \prod_{i=1}^N \langle \Phi_i(Z) | \exp \{ -i\varepsilon H(t_i) \} | \Phi_{i-1}(Z) \rangle \\ &= \int [d\mu(Z(t))] \exp \{ iS[Z(t)] \}, \end{aligned} \quad (128)$$

where

$$S[Z(t)] = \int_{t_0}^{t_f} dt \left\{ \langle \Phi(Z(t)) | i \frac{d}{dt} | \Phi(Z(t)) \rangle - \langle \Phi(Z(t)) | H(t) | \Phi(Z(t)) \rangle \right\} \quad (129)$$

is an effective action in the generalized coherent state representation. This path integral is defined over the coset space  $\mathbf{G}/\mathbf{H}$ . The effective action contains two terms. The second term is the matrix element of Hamiltonian operator in the coherent states, which determines the static properties of the classical Hamiltonian. The first term is pure geometric, and it is indeed a Berry phase [47, 48] that describes quantum fluctuations, and also determines the time-evolution of the system,

$$\omega[\mathbf{G}/\mathbf{H}] = \int_{\Gamma \in \mathbf{G}/\mathbf{H}} \langle \Phi(Z) | d | \Phi(Z) \rangle = \int_{\Gamma \in \mathbf{G}/\mathbf{H}} \mathcal{A} \cdot d\hat{\Omega}, \quad (130)$$

where  $\mathcal{A}$  is a gauge vector potential defined over the coset space  $\mathbf{G}/\mathbf{H}$ , and  $\hat{\Omega}$  is a unit vector in  $\mathbf{G}/\mathbf{H}$ . One can then define the gauge connection,

$$F \equiv \langle d\Phi(Z) | d\Phi(Z) \rangle = \sum_{\alpha\alpha'} \omega_{\alpha\alpha'} dZ_\alpha \wedge dZ_{\alpha'} \quad (131)$$

and  $\omega_{\alpha\alpha'}$  is the Berry curvatures:

$$\omega_{\alpha\alpha'} = \left\langle \frac{\partial \Phi(Z)}{\partial Z_\alpha} \left| \frac{\partial \Phi(Z)}{\partial Z_{\alpha'}} \right. \right\rangle - \left\langle \frac{\partial \Phi(Z)}{\partial Z_{\alpha'}} \left| \frac{\partial \Phi(Z)}{\partial Z_\alpha} \right. \right\rangle. \quad (132)$$

When the rank of  $\mathbf{G}$  is larger than one, the associated gauge potential  $\mathcal{A}$  is non-abelian with gauge group  $\leq \mathbf{H}$ . From the above generalized coherent state path integral, one can study the so-called geometric quantization [50, 51] and classical gauge equations of motion in quantum mechanics [52, 53].

This path integral formulation and the associated gauge potentials have potential applications in condensed matter physics and particle physics. This is because classical semisimple Lie groups can be generated by bilinear operators of bosonic and fermionic creation and annihilation operators. The bilinear operators describe the basic collective excitations in strongly correlated or strongly coupled systems. Therefore, the above formalism can be applied directly to various realistic physical problems.

Specifically, the  $SU(n)$  group can be generated by the particle-hole pairs:  $\{a_i^\dagger a_j, 1 \leq i, j \leq n\}$ , and the corresponding generalized coherent state is given by

$$|\{Z_{ij}\}\rangle = \mathcal{N}(Z) \exp \left\{ \sum_{ij} Z_{ij} a_i^\dagger a_j \right\} |m\rangle. \quad (133)$$

where  $a_i^\dagger, a_i$  can be either bosonic or fermionic creation and annihilation operators, and  $|m\rangle$  contains  $m$  particles in the lowest states,  $m < n$ . For bosonic system, the coherent states are defined on the coset space  $\{Z_{ij}\} \in SU(n)/SU(n-1) \times U(1)$ . For fermionic space, the coset space  $\{Z_{ij}\}$  is  $SU(n)/SU(n-m) \times U(m)$ . The spin coherent state of the Heisenberg model discussed in the last section is a special case where the spin operators take the form:

$$\vec{S}_i = \frac{1}{2} \sum_{\alpha\beta} a_{i\alpha}^\dagger \vec{\sigma}_{\alpha\beta} a_{i\beta}, \quad (134)$$

$\sigma$  is the Pauli matrix, and  $\alpha, \beta$  denote the spin index of electrons. Let  $Z_{ii} = \tan \frac{\theta}{2} e^{-i\varphi}$ , the spin coherent state (103) can be reduced to the form of (133) with  $Z_{ij} = 0$  for  $i \neq j$  and  $\mathcal{N}(Z) = 1/(1+|Z_{ii}|^2)^s$ . Correspondingly, the geometrical space  $SU(n)/SU(n-m) \times U(m)$  is reduced to  $\{Z_{ii}\} \in \prod_i \otimes SU_i(2)/U_i(1)$ .

The  $Sp(2n+1)$  group can be realized by bosonic particle-particle and particle-hole pairs:  $\{a_i^\dagger, a_i, a_i^\dagger a_j^\dagger, a_i a_j, a_i^\dagger a_j - \frac{1}{2}\delta_{ij}\}$ . The generalized coherent state of  $Sp(2n+1)$  is given by

$$|\{z_i, Z_{ij}\}\rangle = \mathcal{N}(Z) \exp \left\{ \sum_i (z_i a_i^\dagger - z_i^* a_i) \right\} \exp \left\{ \sum_{ij} \frac{1}{2} Z_{ij} a_i^\dagger a_j^\dagger \right\} |0\rangle. \quad (135)$$

The corresponding coset space  $\{z_i, Z_{ij}\}$  is  $SP(2n+1)/U(n)$ . The squeezed coherent states discussed in Sec. V are only special cases of the above coherent state.

The  $SO(2n)$  group can be generated by fermionic particle-particle and particle-hole pairs  $\{c_i^\dagger c_j^\dagger, c_i c_j, c_i^\dagger c_j - \frac{1}{2}\delta_{ij}\}$ . Similarly, one can write down the most general coherent state for  $SO(2n)$ :

$$|\{Z_{ij}\}\rangle = \mathcal{N}(Z) \exp \left\{ \sum_{ij} Z_{ij} c_i^\dagger c_j^\dagger \right\} |0\rangle, \quad (136)$$

and its geometrical space is the coset space  $\{Z_{ij}\} \in SO(2n)/U(n)$ . A typical example of the above coherent states is the BCS superconducting state in which only special fermionic pairs, i.e., Cooper pairs  $c_{k\uparrow}^\dagger c_{-k\downarrow}^\dagger$ , are considered [54]:

$$|\text{BCS}\rangle = \frac{1}{\sqrt{1+|h_k|^2}} \exp \left\{ \sum_k h_k c_{k\uparrow}^\dagger c_{-k\downarrow}^\dagger \right\} |0\rangle. \quad (137)$$

Since  $\{c_{k\uparrow}^\dagger c_{-k\downarrow}^\dagger, c_{-k\downarrow} c_{k\uparrow}, c_{k\uparrow}^\dagger c_{k\uparrow} + c_{-k\downarrow}^\dagger c_{-k\downarrow} - 1\}$  span a  $su(2)$  algebra, the geometrical space of the above BCS states is indeed the same as the spin coherent state in Heisenberg model, i.e.,  $\{h_k\} \in \prod_k \otimes SU_k(2)/U_k(1)$ . Therefore, the BCS state carries a  $U(1)$  gauge degree of freedom. But physically, the superconductivity is very different from the ferro and antiferro-magnetism. This is because the Heisenberg model has a global spin rotational symmetry, while the BCS Hamiltonian only has a global  $U(1)$  symmetry. In the Heisenberg model, the spontaneously breaking of spin rotational symmetry leads to the spin-wave excitations which can be described by the Non-Linear Sigma Model derived from the spin coherent state path integral, as we have discussed in the previous section. In the BCS theory, the spontaneously breaking of the  $U(1)$  symmetry for the pairing coherence gives pair excitations which can be described by Ginzberg-Landau theory. The Ginzberg-Landau theory should also be derivable from the path integral of BCS coherent states.

The above general fermionic pairing coherent states can also be applies to systems other than the conventional BCS superconductivity. For example, if I take the triplet pairs:

$$\vec{T}^+(k) = \frac{1}{2} \sum_{\alpha\beta} c_{k\alpha}^\dagger (i\vec{\sigma}_2)_{\alpha\beta} c_{-k\beta}^\dagger, \quad \vec{T}(k) = \frac{1}{2} \sum_{\alpha\beta} c_{-k\alpha} (-i\sigma_2 \vec{\sigma})_{\alpha\beta} c_{k\beta} \quad (138)$$

together with the charge and spin operators:

$$Q(k) = \frac{1}{2} \sum_{\alpha} (c_{k\alpha}^\dagger c_{k\alpha} + c_{-k\alpha}^\dagger c_{-k\alpha}) - 1, \quad \vec{S}(p) = \frac{1}{2} \sum_{\alpha\beta} (c_{k\alpha}^\dagger \vec{\sigma}_{\alpha\beta} c_{k\beta} + c_{-k\alpha}^\dagger \vec{\sigma}_{\alpha\beta} c_{-k\beta}), \quad (139)$$

which generates a  $SO(5)$  group, I can construct a generalized coherent state for the triplet pairing for superfluid  $^3\text{He}$ :

$$|\text{SF}\rangle = \mathcal{N}(Z) \exp \sum \{ \vec{Z}_k \cdot \vec{T}^+(k) \} |0\rangle. \quad (140)$$

Its coset space is  $\{\vec{Z}_k\} \in \prod_k \otimes SO_k(5)/U_k(2)$ . This  $SO(5)$  coherent state carries a non-abelian  $SU(2)$  gauge. One can use this  $SO(5)$  generalized coherent state to study non-abelian gauge fields

and low energy effective theory for superfluid  $^3\text{He}$  atoms [55]. Recently, I also constructed a generalized pairing state to include the singlet and triplet pairs,

$$\begin{aligned} |ZW\rangle = \mathcal{N}(Z) \prod_{\mathbf{k}}' \exp \{ & Z_1(\mathbf{k}) c_{\mathbf{k}\uparrow}^\dagger c_{-\mathbf{k}\downarrow}^\dagger + Z_2(\mathbf{k}) c_{\mathbf{k}+\mathbf{Q}\uparrow}^\dagger c_{-\mathbf{k}+\mathbf{Q}\downarrow}^\dagger \\ & + Z_3(\mathbf{k}) c_{\mathbf{k}\uparrow}^\dagger c_{-\mathbf{k}+\mathbf{Q}\downarrow}^\dagger + Z_4(\mathbf{k}) c_{\mathbf{k}\downarrow}^\dagger c_{-\mathbf{k}+\mathbf{Q}\uparrow}^\dagger \\ & + Z_5(\mathbf{k}) c_{\mathbf{k}\uparrow}^\dagger c_{-\mathbf{k}+\mathbf{Q}\uparrow}^\dagger + Z_6(\mathbf{k}) c_{\mathbf{k}\downarrow}^\dagger c_{-\mathbf{k}+\mathbf{Q}\downarrow}^\dagger \} |0\rangle, \end{aligned} \quad (141)$$

for the study of high  $T_c$  superconductivity and the close proximity between the Mott insulating antiferromagnetic order and  $d$ -wave superconducting order in cuprates[64]. Here the coset space is  $\prod_{\mathbf{k}} \otimes \text{SO}_k(8)/\text{U}_k(4)$ . Under the constraint of non-double occupied sites, possible gauge group contains in the above coherent pairing states may be  $\text{SU}(2) \times \text{U}(1)$  or a larger one up to  $\text{U}(4)$ . This may open a new window for the study of the dynamical mechanism of high  $T_c$  superconductivity.

If one takes the continuum limit in the coordinate space and lets  $t_0 \rightarrow -\infty$ ,  $t_f \rightarrow \infty$ , then the path integral based on the generalized coherent states can be expressed as

$$G = \int [d\mu(\hat{\Omega}(x))] \exp \left\{ i \int d^{d+1}x \left\{ \mathcal{A}(x) \cdot \hat{\Omega}(x) - \mathcal{H}[\hat{\Omega}(x)] \right\} \right\}, \quad (142)$$

where  $x$  is a coordinate in the  $d+1$  dimensional Minkowski space. If the Hamiltonian has a symmetry  $\mathbf{S} \subset \mathbf{G}$ , and the static classical ground states (which can be obtained by minimizing  $\mathcal{H}$  with respect to the coherent state parameters) spontaneously breaks this symmetry, then one can use the saddle-point expansion to derive a general Non-Linear Sigma Model defined on  $\mathbf{G}/\mathbf{H}$ ,

$$G = \int [d\mu(\hat{\Omega}(x))] \exp \left\{ i \int d^{d+1}x \left\{ \frac{1}{2g} \partial_\mu \hat{\Omega}(x) \cdot \partial^\mu \hat{\Omega}(x) + \dots + \Theta_{\text{Top}} \right\} \right\}, \quad (143)$$

to describe the low energy physics in the long-wave length limit, where  $\{\dots\}$  denotes the higher order derivatives in the Non-Linear Sigma Model, and  $\Theta_{\text{Top}}$  is a topological phase, corresponding to a Wess-Zumino-Witten topological term [61, 59, 60] that is induced by the gauge degrees of freedom contained in the generalized coherent states and/or a Chern-Simons term of topological gauge fields over the coset space  $\mathbf{G}/\mathbf{H}$  [68]. There may exist many potential applications of such a Non-linear Sigma Model in real physical problems, such as quantum Hall effect [60, 62], the high  $T_c$  superconductivity [63, 64], the disorder systems [65] in condensed matter physics. It is also possible to apply such theory to quantum chromodynamics in particle physics when quantum chromodynamics is formulated in lattices[66, 67]), and quantum gravity [68], etc.

## 8 Summation

In summation, as I have emphasized throughout the article, coherent states possess three unique properties that are fundamental to field theory. The property of coherent behavior uniquely describes the processes involving infinite number of virtual particles. The coherent excitations obtained from coherent states also give the essential physical picture of long-range orders induced by strong correlations. The property of overcompleteness provides a reformulation and generalization of the functional integral in field theory, in which quantum fluctuations of composite operators are included as new low energy dynamical field variables. One may thereby be able to determine the dynamical degrees of freedom in different energy scales and to derive the corresponding effective theory. The property of topologically nontrivial geometrical structure in generalized coherent states allows one to explore the origin of gauge fields and associated gauge degrees of freedom. In this article, I have not touched the recently development of coherent states in terms of superalgebras and quantum groups. These topics may also be very important in the modern development of field theory, such as in supersymmetry, superstring and conformal field theory. Nevertheless, in my personal opinions, understanding the origin as well as the nature of gauge degrees of freedom in physics is perhaps the most fundamental problem in field theory.

## References

- [1] E. Schrödinger, *Naturwissenschaften*, **14**, 644 (1926).
- [2] T. D. Lee, F. E. Low and D. Pines, *Phys. Rev.* **90**, 297 (1953)
- [3] P. W. Anderson, *Phys. Rev.* **110**, 827 (1958).
- [4] J. G. Valatin and D. Butler, *Nuovo Cimento*, **10**, 37 (1958).
- [5] J. Schwinger, *Proc. Nat. Acad. Sci.* **37**, 452, 455 (1951); *Phys. Rev.* **92**, 1283 (1953).
- [6] J. R. Klauder, *Ann. Phys.* **11**, 123 (1960).
- [7] R. J. Glauber, *Phys. Rev.* **130**, 2529 (1963); **131**, 2766 (1963).
- [8] E. C. G. Sudarshan, *Phys. Rev. Lett.* **10**, 277 (1963).
- [9] A. M. Perelomov, *Commun. Math. Phys.* **26**, 222 (1972).
- [10] R. Gilmore, *Rev. Mex. de Fisica*, **23**, 143 (1972).
- [11] J. R. Klauder and B-S. Skagerstam, "Coherent States, applications in physics and mathematical physics" (World Scientific, Singapore, 1985).
- [12] A. M. Perelomov, *Generalized Coherent States and Their Applications*, (Spring-Verlag, Berlin, 1986).
- [13] W. M. Zhang, D. H. Feng and R. Gilmore, *Rev. Mod. Phys.* **62**, 867 (1990).
- [14] B-S. Skagerstam, in the Proc. "Coherent States: Past, Present and Future", Ed. by D. H. Feng, J. R. Klauder and M. R. Strayer (World Scientific, Singapore, 1994).
- [15] J. R. Klauder and E. C. G. Sudarshan, "Fundamentals of Quantum Optics" (W. A. Benjamin, New York, 1968).
- [16] V. Chung, *Phys. Rev.* **140**, B1110 (1965).
- [17] M. Greco and G. Rossi, *Nuovo Cimento*, **50**, 167 (1967).
- [18] T. W. B. Kibble, *J. Math. Phys.* **9**, 315 (1968); *Phys. Rev.* **173**, 1527; *ibid.* **174**, 1882; *ibid.* **175**, 1624 (1968).
- [19] P. P. Kulish and L. D. Faddeev, *Theor. Math. Phys.* **4**, 745 (1970).
- [20] D. Zwanziger, *Phys. Rev.* **D11**, 3481; 3504 (1975).
- [21] N. Papanicolaou, *Phys. Rep.* **24**, 229 (1976).
- [22] D. R. Yennie, S. C. Frautschi and H. Suura, *Ann. Phys. (N. Y.)* **13**, 379 (1961).
- [23] S. Weinberg, *Phys. Rev.* **140**, B516 (1965).
- [24] K. E. Eriksson, N. Mukunda and B. S. Skagerstam, *Phys. Rev.* **D24**, 2615 (1982).
- [25] T. D. Lee and M. Nauenberg, *Phys. Rev.* **133**, B1549 (1964).
- [26] G. Sterman, "Introduction to Quantum Field Theory" (Cambridge, 1993).
- [27] R. P. Feynman, *Rev. Mod. Phys.* **20**, 367 (1948).
- [28] R. P. Feynman and A. R. Hibbs, "Quantum Mechanics and Path Integrals" (McGraw-Hill, New York, 1965).

- [29] L. D. Feddeev, in "Methods in Field Theory", Eds. R. Balian and J. Zinn-Justin (North-Holland/World Scientific, Amsterdam, 1976)
- [30] J. R. Klauder, Phys. Rev. **D19**, 2349 (1979).
- [31] Y. Ohnuki and T. Kashiwa, Prog. Theor. Phys. **60**, 548 (1978).
- [32] E. Aders and B. W. Lee, Phys. Rep. **9C**, 1 (1973).
- [33] L. D. Feddeev and A. A. Slavnov, "Gauge Fields - Introduction to Quantum Theory" (W. A. Benjamin, New York, 1980).
- [34] T. T. Wu and C. N. Yang, Phys. Rev. **D12**, 3845 (1975).
- [35] Y. P. Yao, Phys. Rev. Lett. **36**, 653 (1976).
- [36] C. Fabre, Phys. Rep. **219**, 215 (1992).
- [37] D. F. Walls, Nature, **306**, 141 (1983).
- [38] R. Jackiw and A. Kerman, Phys. Lett. **A71**, 158 (1979).
- [39] F. W. Cummings and J. R. Johnston, Phys. Rev. **151**, 105 (1966).
- [40] Y. Tsue and Y. Fujiwara, Prog. Theor. Phys. **86**, 469 (1991).
- [41] K. G. Wilson and J. Kogut, Phys. Rep. **12C**, 75 (1974).
- [42] F. T. Arecchi, E. Couttens, R. Gilmore, and H. Thomas, Phys. Rev. **A6**, 2211 (1972).
- [43] M. V. Berry, Proc. R. Soc. London, Ser. **A392**, 45 (1984).
- [44] D. C. Mattis, "Theory of Magnetism" (Spring-Verlag, 1988)
- [45] R. M. White, "Quantum Theory of Magnetism" (Spring-Verlag, 1987).
- [46] F. D. M. Haldane, Phys. Lett. **A93**, 464 (1983); Phys. Rev. Lett. **50**, 1153 (1983).
- [47] B. Simon, Phys. Rev. Lett. **51**, 2167 (1983).
- [48] F. Wilczek and A. Zee, Phys. Rev. Lett. **52**, 2111 (1984).
- [49] A. Shapere and F. Wilczek, "Geometric Phases in Physics" (World Scientific, Singapore, 1989).
- [50] N. M. J. Woodhouse, "Geometric Quantization" 2nd Ed. (Oxford, 1992).
- [51] A. Aleksev, L. Feddeev and S. Shatashvili, J. Geom. Phys. **3**, 391 (1989).
- [52] S. K. Wong, Nuovo Cimento, **A65**, 689 (1970).
- [53] A. P. Balachandran, G. Marmo, S.-B. Skagerstam and A. Stern, "Classical Topology and Quantum States" (World Scientific, Singapore, 1991).
- [54] J. Bardeen, L. N. Cooper and J. R. Schrieffer, Phys. Rev. **108**, 1175 (1957).
- [55] A. J. Leggett, Rev. Mod. Phys. **47**, 331 (1975).
- [56] G. E. Volvik, "Exotic Properties of Superfluid  $^3\text{He}$ " (World Scientific, Singapore, 1992).
- [57] W. M. Zhang, cond-mat/9907287.
- [58] E. Witten, Commun. Math. Phys. **92**, 455 (1984); *ibid.* **137**, 29 (1991).
- [59] P. B. Wiegman, Phys. Rev. Lett. **60**, 821 (1988).

- [60] M. Stone, Phys. Rev. Lett. **63**, 731 (1989); Nucl. Phys. **B227**, 399 (1989).
- [61] E. Witten, Commun. Math. Phys. **121**, 351 (1989).
- [62] M. R. Zirnbauer, hep-th/9905054, and references therein.
- [63] S. C. Zhang, Science, **275**, 1196 (1997).
- [64] E. Demler and S. C. Zhang, Ann. Phys. **271**, 83 (1999).
- [65] K. Efetov, "Supersymmetry in Disorder and Chaos" (Cambridge, 1997).
- [66] K. G. Wilson, Phys. Rev. **D10**, 2445 (1974); and In "New Phenomena in Subnuclear Physics", Ed. A. Zichichi (Plenum Press, New York, 1977).
- [67] S. R. Sharpe, hep-lat/9811006.
- [68] E. Witten, Nucl. Phys. **B323**, 113 (1989); Phys. Rev. **D44**, 314 (1991).

# 12. Pancharatnam, Bargmann and Berry Phases - A Retrospective

N. Mukunda <sup>†</sup>

Centre for Theoretical Studies and Department of Physics,  
Indian Institute of Science, Bangalore 560 012, India

## 1 Introduction

Berry's discovery in 1983-84[1] of "quantum adiabatic anholonomy" has been the starting point and inspiration for an enormous amount of work exploring its properties, generalisations, reformulations and applications[2]. Now known as the geometric phase, Berry himself has characterised it as something that helped clear up "a corner of quantum mechanics that was for a long time dusty and obscure"[3]. It seems fair to say that this concept has achieved much more, and has shown astonishing conceptual depth and unifying power.

While various generalisations of the original formulation came soon after Berry's initial work, it was also discovered in due course that there were several precursors to his ideas. In retrospect, two of these seem specially significant - work by S.Pancharatnam in 1956 in the arena of classical polarization optics[4], and by V.Bargmann in 1964 in the context of Wigner's theorem on the representation of symmetry operations in quantum mechanics[5]. In both cases, phases of complex quantities played important roles, and today it seems evident that these and the later geometric phase all belong to the same circle of ideas.

The purpose of this paper is to describe these developments in such a way that one sees in proper perspective the close connections between the ideas of Pancharatnam, Bargmann and Berry and how they dovetail into one another. The emphasis, admittedly selective, will be on the theoretical aspects. In section 2 we briefly review the original derivation by Berry of the geometric phase, spelling out the assumptions made on the way. This is then followed by an account of the generalisations due to Aharonov and Anandan on the one hand[6], and by Samuel and Bhandari on the other[7]. These relax to a great extent the conditions under which geometric phases can be defined. Section 3 carries these processes further and exhibits these phases as a part of quantum kinematics, namely as reflecting the presence of a complex linear vector space structure as state space[8]. One aspect of the close connection between group actions and geometric phases comes through in this manner, namely the phase is seen to be the simplest invariant expression under certain well defined groups of transformations acting on curves in Hilbert space. Sections 4 and 5 are devoted to explanations of the links to the two precursors mentioned above, namely to the work of Bargmann and of Pancharatnam respectively. These sets of ideas intersect in diverse ways and this is brought out in some detail. Section 6 returns to the theme of the connection between groups and geometric phases but in a new manner. The focus now is on the structure of geometric phases which arise from unitary Lie group representations[9]. While this could be viewed as a particular case of the general theory, it merits separate discussion as there is a rich interplay between Lie algebraic and differential geometric structures and these are reflected in the final geometric phase formula. Section 7 contains concluding remarks.

---

\*Email: nmukunda@cts.iisc.ernet.in

<sup>†</sup>Honorary Professor, Jawaharlal Nehru Centre for Advanced Scientific Research, Jakkur, Bangalore 560064

## 2 The adiabatic geometric phase and its generalisations

Consider a quantum mechanical system with state vector  $\psi(t)$  at time  $t$  and hermitian time dependent Hamiltonian operator  $\hat{H}(t)$ , evolving according to the Schrödinger equation

$$i \frac{d}{dt} \psi(t) = \hat{H}(t) \psi(t). \quad (1)$$

(Here for simplicity we have set Planck's constant  $\hbar = 1$ ). Assume that the time dependence of  $\hat{H}(t)$  is mild and that we are able to apply the adiabatic theorem of quantum mechanics. To this end, let  $u(t)$  be a normalised nondegenerate instantaneous eigenvector of  $\hat{H}(t)$  with (possibly time dependent) eigenvalue  $E(t)$ :

$$\begin{aligned} \hat{H}(t)u(t) &= E(t)u(t), \\ (u(t), u(t)) &= \|u(t)\|^2 = 1 \end{aligned} \quad (2)$$

Then, provided the overall phase of  $u(t)$  at each time is adjusted so that

$$\left( u(t), \frac{du(t)}{dt} \right) = 0, \quad (3)$$

we can construct an approximate solution  $\psi_{ad}(t)$  of eqn.(2.1) by setting

$$\begin{aligned} \psi_{ad}(t) &= \exp \left( -i \int_0^t E(t') dt' \right) u(t), \\ i \frac{d}{dt} \psi_{ad}(t) &\simeq \hat{H}(t) \psi_{ad}(t) \end{aligned} \quad (4)$$

The explicit time dependence in  $\hat{H}(t)$  can be thought of as due to the quantum system being placed in a slowly changing (classical) environment. Let the environment now be supposed to be cyclic, in the sense that at time  $t = T$  it returns to its condition at, say,  $t = 0$ . This is reflected in the operator  $\hat{H}(t)$  similarly being cyclic:

$$\hat{H}(T) = \hat{H}(0) \quad (5)$$

One can now ask for the relation between  $\psi_{ad}(T)$  and  $\psi_{ad}(0)$ : is it also similarly cyclic? That the physical state is cyclic (provided the nondegeneracy condition is obeyed throughout  $t = 0$  to  $t = T$ ) is well known. This means that  $\psi_{ad}(T)$  can differ from  $\psi_{ad}(0)$  at most by a phase:

$$\begin{aligned} \psi_{ad}(T) &\simeq e^{i\varphi_{\text{tot}}} \psi_{ad}(0), \\ \varphi_{\text{tot}} &= \text{total phase} \end{aligned} \quad (6)$$

One contribution to this total phase is evident from the adiabatic theorem and the definition (2.4) of  $\psi_{ad}(t)$ : this is the dynamical phase

$$\varphi_{\text{dyn}} = - \int_0^T dt E(t). \quad (7)$$

Berry's surprising discovery was that there is another geometric contribution  $\varphi_{\text{geom}}$  to  $\varphi_{\text{tot}}$ , namely that

$$\varphi_{\text{tot}} = \varphi_{\text{dyn}} + \varphi_{\text{geom}}, \quad (8)$$

where  $\varphi_{\text{geom}}$  is determined in a global way by, or belongs to, the entire path traced out in state space by  $\psi(t)$  from  $t = 0$  to  $t = T$ . The important point is that this piece  $\varphi_{\text{geom}}$  is robust in the



sense that it is unaffected by the various phase choices or freedoms available in the problem. This  $\varphi_{\text{geom}}$  is the geometric phase. One can trace it to the fact that if  $u(t)$  obeys (2.3) throughout, then  $u(T)$  and  $u(0)$  in general differ by a phase.

Thus it is clear that the original discovery of the geometric phase was in the context of unitary, adiabatic, cyclic Schrödinger evolution in quantum mechanics. It may also be mentioned that Berry made essential use of the classical parameter space which accounts for the explicit time dependence of  $\hat{H}(t)$ ; and the explicit expressions that were developed for  $\varphi_{\text{geom}}$  were also written in terms of line and surface integrals over this space.

In gradual degrees the original restrictions or conditions under which the geometric phase was defined were relaxed or removed. In the work of Aharonov and Anandan[6] it was shown that one can give up the adiabatic condition and still define the geometric phase. As long as one has a solution of the time dependent Schrödinger equation (2.1), and the solution is cyclic in the sense that

$$\psi(T) = e^{i\varphi_{\text{tot}}} \psi(0) \quad (9)$$

whether or not the Hamiltonian obeys the cyclic condition (2.5), one can define the geometric phase by

$$\begin{aligned} \varphi_{\text{geom}} &= \varphi_{\text{tot}} - \varphi_{\text{dyn}}, \\ \varphi_{\text{dyn}} &= - \int_0^T dt (\psi(t), \hat{H}(t) \psi(t)). \end{aligned} \quad (10)$$

The main point is that the definition (2.7) of the dynamical phase has been generalised beyond the confines of the adiabatic theorem; the latter is of course recovered as a special case. Moreover,  $\varphi_{\text{geom}}$  clearly belongs to the particular cyclic solution  $\psi(t)$ ,  $0 \leq t \leq T$ , to the Schrödinger equation that one has in hand. Aharonov and Anandan also showed that the geometric phase is intrinsically something defined for cyclic evolutions in the quantum mechanical state space or ray space associated with the Hilbert space, rather than in the classical parameter space. But here again one can easily recover the original framework of Berry as a particular case, and that is quite often useful.

The next important extension, due to Samuel and Bhandari[7], showed that the geometric phase could be defined for noncyclic and even for nonunitary evolutions. As for the former, the key idea was to introduce and exploit the properties of geodesics in quantum mechanical ray and Hilbert spaces. Thus a case of noncyclic evolution was converted to cyclic evolution by joining the end points via a geodesic, and then using the earlier definition for  $\varphi_{\text{geom}}$  one has a definition valid for noncyclic cases. The case of nonunitary evolution corresponds to working with a Schrödinger equation in which the Hamiltonian may be nonhermitian.

All these extensions demonstrate the robustness of the geometric phase concept in that it seems to require very limited formal machinery to be able to identify and define this phase. This feature will be reinforced when we describe the kinematic approach in the next section. On the experimental side, many examples of the geometric phase have been demonstrated. Purely for illustration we may mention that the early experiment of Chiao et al[10] showed the presence of this phase in a classical optical situation; the experiment of Bhandari et al[11] was in the original Pancharatnam framework of polarization optics; while that of Simon et al[13] showed how a time-dependent geometric phase could be converted into a frequency shift capable of fine-tuning laser beams.

### 3 The geometric phase seen via quantum kinematics

Now we show that one can go one step further in relaxing the conditions necessary for the definition of the geometric phase - even the Hamiltonian operator and the Schrödinger equation can be given

up, thus showing that this concept depends solely on the existence of a complex linear vector space structure. This may aptly be called the quantum kinematic approach to the geometric phase[8].

Let  $\mathcal{H}$  be the Hilbert space describing the pure states of some quantum system - it may be of finite or infinite dimension. For the most part all the constructions to follow involve the set of unit vectors in  $\mathcal{H}$ , namely the unit sphere in  $\mathcal{H}$ . The basic mathematical objects we deal with are continuous, almost everywhere differentiable generally open curves  $\mathcal{C} \subset \mathcal{H}$  parametrised by a monotonically increasing parameter  $s$ :

$$\mathcal{C} = \{\psi(s) \in \mathcal{H} \mid \|\psi(s)\| = 1, s_1 \leq s \leq s_2\} \quad (1)$$

The ray space  $\mathcal{R}$  associated to  $\mathcal{H}$  consists of the set of all projection operators  $\hat{\rho}$  onto unit vectors in  $\mathcal{H}$ :

$$\begin{aligned} \mathcal{R} &= \{\hat{\rho} = \psi\psi^\dagger \mid \psi \in \mathcal{H}, \|\psi\| = 1\} \\ &= \{\hat{\rho} = \text{operator on } \mathcal{H} \mid \hat{\rho}^\dagger = \hat{\rho}, \hat{\rho} \geq 0, \text{Tr } \hat{\rho} = 1, \hat{\rho}^2 = \hat{\rho},\} \end{aligned} \quad (2)$$

Every curve  $\mathcal{C} \subset \mathcal{H}$  described as in eqn.(3.1) possesses a ray space image  $C \subset \mathcal{R}$ , namely a continuous almost everywhere differentiable parametrised curve of normalised pure state density matrices. Denoting by  $\pi$  the natural projection  $\mathcal{H} \rightarrow \mathcal{R}$  (as in  $\mathcal{H}$ , so with  $\mathcal{R}$  we are mainly concerned with normalised  $\hat{\rho}$ ), we have

$$C = \pi[\mathcal{C}] = \{\hat{\rho}(s) \mid \hat{\rho}(s) = \pi(\psi(s)) = \psi(s)\psi(s)^\dagger, s_1 \leq s \leq s_2\} \subset \mathcal{R} \quad (3)$$

For given  $C \subset \mathcal{R}$ , any  $\mathcal{C} \subset \mathcal{H}$  such that  $\pi[\mathcal{C}] = C$  is a lift of  $C$ .

We now define two groups of transformations on such curves  $\mathcal{C}$  mapping them onto similar curves  $\mathcal{C}'$ : the group of local phase changes, and the group of continuous monotonic reparametrisations. They act as follows: (a) Local phase changes:

$$\mathcal{C} = \{\psi(s) \mid s_1 \leq s \leq s_2\} \longrightarrow \mathcal{C}' = \{\psi'(s) = e^{i\alpha(s)}\psi(s) \mid \alpha(s) \text{ real}, s_1 \leq s \leq s_2\}; \quad (4)$$

(b) Reparametrisations:

$$\begin{aligned} \mathcal{C} = \{\psi(s) \mid s_1 \leq s \leq s_2\} \longrightarrow \mathcal{C}' = \{\psi'(s') = \psi(s) \mid s' = f(s) = \text{real monotonic}, \\ s'_1 = f(s_1) \leq s' \leq s'_2 = f(s_2)\} \end{aligned} \quad (5)$$

We see that under transformations (a), the ray space image  $C$  (and its parametrisation) are left intact; while in transformations (b) the points comprising  $\mathcal{C}$  (and similarly  $\mathcal{C}'$ ) are left intact but only their parametrisation or labelling gets changed.

Now we ask for the simplest (real valued) expression or functional of  $\mathcal{C}$  that is invariant under both groups of transformations defined above. A little reflection shows that the following expression, tentatively denoted by  $\varphi_g$ , has these properties:

$$\begin{aligned} \varphi_g &= \varphi_P - \varphi_{\text{dyn}}, \\ \varphi_P &= \arg(\psi(s_1), \psi(s_2)), \\ \varphi_{\text{dyn}} &= \text{Im} \int_{s_1}^{s_2} ds \left( \psi(s), \frac{d\psi(s)}{ds} \right). \end{aligned} \quad (6)$$

Several remarks are in order at this stage. Since  $\psi(s)$  is normalised, the inner product  $\left(\psi(s), \frac{d\psi(s)}{ds}\right)$  is in any case pure imaginary. While the quantity  $\varphi_{\text{dyn}}$  is quite well defined, the (nonlocal) quantity  $\varphi_P$ , and so  $\varphi_g$ , are both defined modulo  $2\pi$ ; this is unavoidable. The two pieces  $\varphi_P, \varphi_{\text{dyn}}$  are individually invariant under the reparametrisation transformations (3.4b); but only their difference is invariant under the phase change transformations (3.4a), so they are tied together in this way. One could have written expressions exactly like the above but referring to only a (connected) subset

of  $\mathcal{C}$ , say running from  $s_3$  to  $s_4$  where  $s_1 < s_3 < s_4 < s_2$ ; but then in a sense the whole of  $\mathcal{C}$  would not have been used in the construction.

It now turns out that the real-valued functional  $\varphi_g$  defined in eqn.(3.5) is the geometric phase associated with the curve  $\mathcal{C}$ . One thus has reduced the prerequisites needed for the definition of this phase really to the bare minimum. Needless to say, all the earlier definitions are easily recovered as special cases; moreover it is clear that the definition of  $\varphi_g$  for an open  $\mathcal{C}$  ("noncyclic evolution") is actually very simple and direct and does not require special considerations. Indeed in this kinematic view, the quantity immediately defined is the noncyclic geometric phase, while the cyclic case is a simple specialisation.

The two invariances leading to  $\varphi_g$  can now be interpreted. Invariance under local phase changes means that  $\varphi_g$  is actually a functional of the ray space image  $C = \pi[\mathcal{C}]$  of  $\mathcal{C}$ , not of  $\mathcal{C}$  itself. Invariance under reparametrisations entitles the use of the term 'geometric'. With these insights, we can rewrite eqn.(3.5) more explicitly, indicating the arguments of each object:

$$\begin{aligned}\varphi_g[C] &= \varphi_P[C] - \varphi_{\text{dyn}}[C], \\ \varphi_P[C] &= \arg(\psi(s_1), \psi(s_2)), \\ \varphi_{\text{dyn}}[C] &= \text{Im} \int_{s_1}^{s_2} \left( \psi(s), \frac{d\psi(s)}{ds} \right) ds.\end{aligned}\tag{7}$$

We can now appreciate: given a ray space curve  $C$ , the calculation of its geometric phase  $\varphi_g[C]$  is facilitated by going to any lift  $\mathcal{C}$  of  $C$ , calculating more easily the two quantities  $\varphi_P[C]$ ,  $\varphi_{\text{dyn}}[C]$ , and then taking their difference - this difference is lift independent and always gives us  $\varphi_g[C]$ .

The subscript  $P$  in  $\varphi_P$  stands for 'Pancharatnam' - the explanation for this will be given in the sequel. Since  $\varphi_g[C]$  is independent of the choice of the lift  $\mathcal{C}$ , it is clear that certain lifts may enjoy special properties. Two of these may be mentioned. 'Pancharatnam lifts' are lifts  $\mathcal{C}$  such that the term  $\varphi_P[C]$  vanishes; then the geometric phase reduces to the dynamical part alone, but there is still a great deal of local phase freedom in the choice of  $\mathcal{C}$ . 'Horizontal lifts' are lifts  $\mathcal{C}$  such that the other term  $\varphi_{\text{dyn}}[C]$  vanishes, so the geometric phase reduces to the Pancharatnam contribution alone. With these lifts the remaining freedom is much more limited.

## 4 Geodesics and the Bargmann invariants

Now we develop the link to the 1964 work of Bargmann[5], and show how that is a precursor to the geometric phase. For this we begin with the concept of geodesics in ray and Hilbert spaces, of course within the families of unit rays and unit vectors.

Given a curve  $\mathcal{C} \subset \mathcal{H}$  with image  $C \subset \mathcal{R}$ , we can define a functional  $L[C]$  which can be interpreted as the length of  $C$ . As with the case of  $\varphi_g[C]$ , it is simplest to compute  $L[C]$  using some lift  $\mathcal{C}$  of  $C$ , and then recognise that because of the built-in invariances we have a functional of  $C$  alone. So, given  $\mathcal{C} = \{\psi(s)\}$ , we define to begin with:

$$\begin{aligned}u(s) &= \frac{d\psi(s)}{ds}, \\ u_{\perp}(s) &= u(s) - \psi(s)(\psi(s), u(s)); \\ \psi'(s) &= e^{i\alpha(s)}\psi(s) \implies u'_{\perp}(s) = e^{i\alpha(s)}u_{\perp}(s)\end{aligned}\tag{8}$$

The component  $u_{\perp}(s)$  of  $u(s)$  is seen to 'transform covariantly', ie., in the same way as  $\psi(s)$  itself, under local phase changes. We then set up  $L[C]$  as:

$$L[C] = \int_{s_1}^{s_2} ds \|u_{\perp}(s)\| = \int_{s_1}^{s_2} ds \{ \|u(s)\|^2 - |(\psi(s), u(s))|^2 \}^{1/2},\tag{9}$$

and regard this as the length of the ray space curve  $C$ . It is clear that it too possesses the invariances of  $\varphi_g[C]$ , namely local phase change and reparametrisation invariances. It is also seen

to be nondegenerate and nonnegative, thus specifying a Riemannian metric on  $\mathcal{R}$ . This is the well known Fubini- Study metric.

We now define geodesics in  $\mathcal{R}$  to be those curves  $C$  which, for given end points  $\hat{\rho}_1, \hat{\rho}_2$ , minimize the length  $L[C]$ :

$$\hat{\rho}_1 = \psi_1 \psi_1^\dagger, \hat{\rho}_2 = \psi_2 \psi_2^\dagger \text{ fixed, } \delta L[C] = 0 \implies C \text{ is a geodesic from } \hat{\rho}_1 \text{ to } \hat{\rho}_2 \quad (10)$$

The corresponding Euler Lagrange differential equations can be easily derived, and exploiting the invariances of  $L[C]$  they can be solved in the most general case[8]. The result of this analysis is that the most general geodesic  $C_{\text{geo}}$  in  $\mathcal{R}$  can be lifted to a (horizontal) curve  $\mathcal{C}_{\text{geo}}$  in  $\mathcal{H}$ , and an (affine) parametrisation can be chosen, so that  $\mathcal{C}_{\text{geo}}$  has the following extremely simple description:

$$\begin{aligned} C_{\text{geo}} \subset \mathcal{R} &\rightarrow \mathcal{C}_{\text{geo}} \subset \mathcal{H}: \\ \mathcal{C}_{\text{geo}} &= \{\psi(s)\}, \\ \psi(s) &= \phi_1 \cos s + \phi_2 \sin s, \\ (\phi_1, \phi_1) &= (\phi_2, \phi_2) = 1, (\phi_1, \phi_2) = 0 \end{aligned} \quad (11)$$

Thus  $\mathcal{C}_{\text{geo}}$  is just an arc of a plane two-dimensional circle in  $\mathcal{H}$ ! (The range of  $s$  can be adjusted so that  $\mathcal{C}_{\text{geo}}$  runs from given  $\hat{\rho}_1$  to given  $\hat{\rho}_2$  in  $\mathcal{R}$ ). We see that for  $\mathcal{C}_{\text{geo}}$ , both  $\varphi_P[\mathcal{C}_{\text{geo}}]$  and  $\varphi_{\text{dyn}}[\mathcal{C}_{\text{geo}}]$  vanish, so the geometric phase for  $C_{\text{geo}}$  also vanishes:

$C_{\text{geo}}$  = geodesic in  $\mathcal{R}$ ,  $\mathcal{C}_{\text{geo}}$  = horizontal affinely parametrised lift :

$$\begin{aligned} \varphi_P[\mathcal{C}_{\text{geo}}] &= \varphi_{\text{dyn}}[\mathcal{C}_{\text{geo}}] = 0, \\ \varphi_g[\mathcal{C}_{\text{geo}}] &= 0. \end{aligned} \quad (12)$$

This - the vanishing of the geometric phase for any geodesic in  $\mathcal{R}$  - is the key result that will lead to the link to the Bargmann invariants.

In the above we worked with specially chosen and convenient lifts  $\mathcal{C}_{\text{geo}}$  of a geodesic  $C_{\text{geo}}$  in  $\mathcal{R}$ . It is convenient to say that any lift  $\mathcal{C}$  of a geodesic in  $\mathcal{R}$  is a geodesic in  $\mathcal{H}$ :

$$\delta L[C] = 0, \pi[C] = C \implies \mathcal{C} \text{ is a geodesic in } \mathcal{H}. \quad (13)$$

Then one has a more flexible result than (4.5):

$$\begin{aligned} \delta L[C] = 0, \pi[C] &= C \implies \\ \varphi_{\text{geo}}[C] &= 0, \\ \varphi_P[C] &= \varphi_{\text{dyn}}[C] \end{aligned} \quad (14)$$

This result will be used below.

Now we define the Bargmann invariants[5]. Let  $\psi_1, \psi_2, \dots, \psi_n$  be any sequence of  $n$  unit vectors in  $\mathcal{H}$ ; they do not have to be mutually orthogonal. Their images in  $\mathcal{R}$  will be denoted as  $\hat{\rho}_1, \hat{\rho}_2, \dots, \hat{\rho}_n$ . The  $n$ -vertex Bargmann invariant is then defined as the product of a sequence of scalar products

$$\Delta_n(\psi_1, \psi_2, \dots, \psi_n) = (\psi_1, \psi_2)(\psi_2, \psi_3) \dots (\psi_{n-1}, \psi_n)(\psi_n, \psi_1) \quad (15)$$

For definiteness, to avoid trivialities, we assume all the factors are nonzero. We see that this expression is invariant under independent phase changes in each of the vectors  $\psi_1, \dots, \psi_n$  involved in its construction. Thus it must really be expressible as a ray space quantity; this is made explicit by writing it as

$$\Delta_n(\psi_1, \psi_2, \dots, \psi_n) = \text{Tr}(\hat{\rho}_1 \hat{\rho}_2 \dots \hat{\rho}_n). \quad (16)$$

For  $n = 1$  or  $n = 2$  we have trivial situations:  $\Delta_1(\psi_1)$  is unity, and  $\Delta_2(\psi_1, \psi_2)$  is real positive. It is from  $n = 3$  onwards that we have something interesting: in general, for  $n \geq 3$ ,  $\Delta_n(\psi_1, \dots, \psi_n)$  is a complex quantity. It can however be easily shown that the essentially new object is  $\Delta_3(\psi_1, \psi_2, \psi_3)$ : for any larger value of  $n$ ,  $\Delta_n$  can be written as the ratio of a product of  $\Delta_3$ 's by a product of  $\Delta_2$ 's. Such an expression however tends to obscure the symmetry of  $\Delta_n(\psi_1, \psi_2, \dots, \psi_n)$  under cyclic permutations of its arguments.

Now comes the link to the geometric phase. The definition (4.8) of the  $n$ -vertex Bargmann invariant requires only the choice of  $n$  'vertices'  $\hat{\rho}_1, \hat{\rho}_2, \dots, \hat{\rho}_n$  in  $\mathcal{R}$  (in a definite sequence), with the vectors  $\psi_1, \psi_2, \dots, \psi_n$  in  $\mathcal{H}$  projecting onto the respective vertices. So far, we have no geometric phases. But to bring them in, we now create a closed figure in  $\mathcal{R}$  by connecting  $\hat{\rho}_1$  to  $\hat{\rho}_2$ ,  $\hat{\rho}_2$  to  $\hat{\rho}_3 \dots$ ,  $\hat{\rho}_n$  to  $\hat{\rho}_1$  by successive geodesic arcs - this results in an  $n$ -sided polygon in  $\mathcal{R}$  with geodesic sides, and one can then ask for the geometric phase for this 'cyclic evolution'. Going back to the basic definition (3.6) and repeatedly exploiting the special property (4.7) of geodesics  $\mathcal{R}$  and their lifts in  $\mathcal{H}$ , we find [8]:

$$\begin{aligned}
 & \varphi_g[n - \text{sided polygon in } \mathcal{R}, \text{ vertices } \hat{\rho}_1, \hat{\rho}_2, \dots, \hat{\rho}_n, \text{ geodesic sides}] \\
 &= \arg(\psi_1, \psi_1) - \varphi_{\text{dyn}}[C_{12} \cup C_{23} \cup C_{34} \dots \cup C_{n-1,n} \cup C_{n,1}] \\
 &= -\varphi_{\text{dyn}}[C_{12}] - \varphi_{\text{dyn}}[C_{23}] - \dots - \varphi_{\text{dyn}}[C_{n,1}] \\
 &= -\varphi_P[C_{12}] - \varphi_P[C_{23}] - \dots - \varphi_P[C_{n,1}] \\
 &= -\arg(\psi_1, \psi_2) - \arg(\psi_2, \psi_3) \dots - \arg(\psi_n, \psi_1) \\
 &= -\arg \Delta_n(\psi_1, \psi_2, \dots, \psi_n).
 \end{aligned} \tag{17}$$

Here  $C_{12}, C_{23}, \dots$  are any lifts of the geodesics from  $\hat{\rho}_1$  to  $\hat{\rho}_2$ ,  $\hat{\rho}_2$  to  $\hat{\rho}_3, \dots$ ; they run in  $\mathcal{H}$  from  $\psi_1$  to  $\psi_2$ ,  $\psi_2$  to  $\psi_3 \dots$ . Equation (4.10) is the connection between Bargmann invariants and geometric phases we are seeking. It is clear that this connection is a direct and simple consequence of the key property (4.5,7) of geodesics - their geometric phases vanish. It should also be clear that for this derivation to be meaningful, it was necessary to regard open curve geometric phases as the primary quantities, and closed curve geometric phases as derived ones.

We may reemphasize the following point. The definition of the Bargmann invariant (4.8) requires only the choice of a sequence of vertices in  $\mathcal{R}$ , there is no need to join these vertices in any way to form a (closed) figure. The latter step has to be taken only to find a connection to geometric phases; when this is done, we see that phases of Bargmann invariants are (apart from a sign) geometric phases.

The point just made, however, motivates us to enlarge this connection. One can ask for the most general kind of curve that may be used to connect the vertices of a Bargmann invariant, such that the phase of the latter is the geometric phase of the resulting closed figure in  $\mathcal{R}$ . This question can be answered and leads to the concept of null phase curves [13]. Geodesics are examples of null phase curves, but the latter are a much wider family than the former. These curves lead to the widest possible generalisation of the connection between Bargmann invariants and geometric phases. The given  $C$  in  $\mathcal{R}$  with lift  $\mathcal{C}$  in  $\mathcal{H}$  is a null phase curve if and only if

$$\begin{aligned}
 \varphi_g[\text{any connected portion of } C] &= 0 \\
 \text{ie. } \varphi_P[\text{any connected portion of } C] &= \varphi_{\text{dyn}}[\text{that portion of } C]
 \end{aligned} \tag{18}$$

Such curves can be nicely characterised at both ray and Hilbert space levels. Then one finds as a generalisation of eqn.(4.10):

$$\arg \Delta_n(\psi_1, \psi_2, \dots, \psi_n) = -\varphi_g[n - \text{sided figure in } \mathcal{R}, \text{ vertices } \hat{\rho}_1, \hat{\rho}_2, \dots, \hat{\rho}_n, \text{ sides as null phase curves}] \tag{19}$$

This generalisation is indeed physically significant, and it leads to important results for instance with respect to coherent states[14], the interpretation of the classical Optical Guoy Phase as a geometric phase[15], etc.

In concluding this Section we may also point out that the link (4.10) can be exploited to recover the original definitions (3.5,6) of the geometric phase by a limiting procedure[8]. All this should convince the reader of the intimate genetic relationship between Bargmann and Berry phases.

## 5 Pancharatnam phases

Now we turn to the third concept in this account, the Pancharatnam phase. Pancharatnam's original work[4] was in the context of classical polarization optics, and he was concerned with the phase relations between two pure states of polarization of a plane electromagnetic wave, such that their interference upon superposition would be maximal. We now explain his ideas in the language of quantum mechanics[16].

Let  $\psi_1$  and  $\psi_2$  be two unit vectors in the Hilbert space  $\mathcal{H}$  of some quantum system. We ask the question - is there a way to define and measure the 'relative phase' of  $\psi_2$  with respect to  $\psi_1$ , and as a particular case can we give a meaning to the expression ' $\psi_1$  and  $\psi_2$  are in phase' ? Pancharatnam's answer, motivated as we said above by the study of superposed pure polarized plane electromagnetic waves, amounts to the following: the vectors  $\psi_1$  and  $\psi_2$  are to be declared as 'in phase' if their Hilbert space inner product  $(\psi_1, \psi_2)$  is real positive; if  $(\psi_1, \psi_2)$  is complex, then its phase can be defined to be the phase of  $\psi_2$  with respect to  $\psi_1$ . (Naturally we assume that  $\psi_1$  and  $\psi_2$  are not orthogonal).

Accepting this definition, we see immediately that the first term  $\varphi_P[C]$  in the definition (3.5,6) of the geometric phase is indeed the phase of the end point  $\psi(s_2)$  of  $C$  with respect to the initial point  $\psi(s_1)$  of  $C$  in the Pancharatnam sense. (This indeed explains our notation for this piece). Calling this a 'Pancharatnam phase' for a moment, we see that it is one of the two ingredients entering into the definition of a general open curve geometric phase. The other is the kinematic version of the Aharonov-Anandan definition of the dynamical phase. So this is one point where the Pancharatnam and the Berry ideas meet.

A natural question now arises: is the Pancharatnam 'in-phase' definition transitive ? In detail, let  $\psi_1, \psi_2, \psi_3$  be any three unit vectors in  $\mathcal{H}$  (no two being orthogonal). Suppose  $\psi_1$  and  $\psi_2$  are 'in phase', and also  $\psi_2$  and  $\psi_3$  are 'in phase', both in the Pancharatnam sense. Then what can we say about the  $\psi_1 - \psi_3$  relationship ? Are they also 'in phase' ? It turns out they are in general not in phase, the Pancharatnam criterion or definition is not transitive, and the mere existence of the Bargmann invariants makes all this immediately obvious! Let us see how this happens.

Given  $\psi_1, \psi_2, \psi_3$ , first form the three- vertex Bargmann invariant

$$\Delta_3(\psi_1, \psi_2, \psi_3) = (\psi_1, \psi_2)(\psi_2, \psi_3)(\psi_3, \psi_1). \quad (20)$$

We are free here to alter the phases of the  $\psi$ 's independently without changing the left hand side. Keeping  $\psi_1$  fixed, let us adjust the overall phase of  $\psi_2$  so that the first factor  $(\psi_1, \psi_2)$  is real positive. This can be done, and it renders  $\psi_1$  and  $\psi_2$  'in phase' in the Pancharatnam sense. Then let us adjust the overall phase of  $\psi_3$  so that the second factor  $(\psi_2, \psi_3)$  also becomes real positive, and  $\psi_2$  and  $\psi_3$  are also 'in phase'. But now we have no more freedom left! If  $\Delta_3(\psi_1, \psi_2, \psi_3)$  was complex to begin with, and this we recall is a ray space statement, then its phase must be reproduced by the third factor  $(\psi_3, \psi_1)$  in the product (5.1). But this phase is known to be a geometric phase. Thus the degree of nontransitivity of the Pancharatnam in - phase concept is expressed via a geometric phase which is just the phase of a three-vertex Bargmann invariant:

$$\psi_1 \text{ and } \psi_2 \text{ 'in phase', } \psi_2 \text{ and } \psi_3 \text{ 'in phase' } \implies$$

$$\text{'phase difference' between } \psi_3 \text{ and } \psi_1 = \arg \Delta_3(\psi_1, \psi_2, \psi_3) \quad (21)$$

This discussion must make it clear how closely interwoven the three streams of ideas really are. For emphasis, at the risk of repetition, let us explicitly bring out these interconnections once again in a series of short statements:

(i) The Pancharatnam relative phase  $\arg(\psi_1, \psi_2)$  between the end points of a curve  $C \subset \mathcal{H}$  is one of the ingredients in the calculation of the geometric phase  $\varphi_g[C]$ .

(ii) Phases of Bargmann invariants are geometric phases of suitably constructed closed ray space figures.

(iii) The very existence of the Bargmann invariants tells us why the Pancharatnam 'in phase' concept is not transitive.

(iv) This extent of nontransitivity is captured in the phase of a Bargmann invariant, and this is a geometric phase.

At this point, let us mention briefly the original calculation of Pancharatnam, expressing it again in the present quantum mechanical language[16]. We have to deal with a two-level quantum system, for which the (complex) dimension of  $\mathcal{H}$  is two. The corresponding ray space is just the Poincare sphere  $S^2$ , which is of real dimension two. If we take three unit vectors  $\psi_1, \psi_2, \psi_3 \in \mathcal{H}$ , their projections to ray space give three real unit vectors  $n_1, n_2, n_3 \in S^2$ . If we connect these points with geodesics or great circle arcs on  $S^2$ , we obtain a spherical triangle on  $S^2$ . Pancharatnam's original result was that if  $\psi_1$  and  $\psi_2$  are 'in phase', and similarly  $\psi_2$  and  $\psi_3$  are 'in phase', then  $\psi_1$  and  $\psi_3$  are 'out of phase', and their 'phase difference' (ie,  $\arg(\psi_3, \psi_1)$ ) is precisely one half of the solid angle subtended at the origin of  $S^2$  by the spherical triangle having  $n_1, n_2, n_3$  for vertices:

$$\dim \mathcal{H} = 2, \psi_1, \psi_2, \psi_3 \in \mathcal{H}, \pi(\psi_{1,2,3}) = n_{1,2,3} \in S^2 :$$

$$\arg \Delta_3(\psi_1, \psi_2, \psi_3) = \frac{1}{2} \Omega,$$

$$\Omega = \text{solid angle of geodesic triangle with vertices } n_1, n_2, n_3. \quad (22)$$

We see many important features of this early result: it is an instance of the relation (4.10) between Bargmann invariants and Berry phases; and as expected the extent to which the amplitudes  $\psi_3$  and  $\psi_1$  are out of phase is determined by a ray space quantity. Lastly, the solid angle on  $S^2$  is an eminently geometric notion.

## 6 Geometric phases from Lie group unitary representations

In many physically interesting situations involving geometric phases, one finds that some unitary representation of some Lie group plays an important role. Thus even though we have not given full details we may mention the following examples: in polarization optics and in the quantum mechanics of a spin 1/2 particle with magnetic moment in a magnetic field, one has to work with the defining representation of the group  $SU(2)$ [1]; geometric phases associated with coherent states involve the Heisenberg Weyl group[14]; squeezing transformations and related geometric phases, as well as the classical optical Guoy phase[15], are concerned with a unitary representation of the metaplectic group  $Mp(2)$ ; three level system geometric phases bring in the defining representation of the group  $SU(3)$ [17]. All these situations suggest that we study in general terms the structures of geometric and dynamical phases produced by unitary Lie group representations. In this Section, we give a brief account of this aspect of geometric phases, just to show the rich interplay between Lie group - Lie algebra structures and geometric phases[9].

Let  $G$  be an  $n$ -dimensional connected Lie group, either compact or noncompact; and let it be unitarily and faithfully represented by operators  $\mathcal{U}(g)$  on the Hilbert space  $\mathcal{H}$  of a quantum system. We do not require irreducibility of this representation. The hermitian generators  $T_r, r = 1, 2, \dots, n$ , of  $\mathcal{U}(g)$  are operators on  $\mathcal{H}$  obeying the commutation relations corresponding to the Lie algebra  $\underline{G}$  of  $G$ :

$$[T_r, T_s] = i f_{rs}{}^t T_t. \quad (23)$$

Here the coefficients  $f_{rs}{}^t$  are the real structure constants of  $G$ . Now choose and keep fixed a fiducial unit vector  $\psi_0 \in \mathcal{H}$ . We are interested in the calculation of dynamical phases for curves

$\mathcal{C} \subset \mathcal{H}$  which begin from  $\psi_0$  and are built up by continuous group action. For this purpose some formal definitions are necessary.

The orbit  $\vartheta(\psi_0)$  of  $\psi_0$  is the set of unit vectors in  $\mathcal{H}$  produced by action of all elements of  $G$  on  $\psi_0$ :

$$\vartheta(\psi_0) = \{\psi(g) = \mathcal{U}(g)\psi_0 | \psi_0 \text{ fixed, } g \in G\} \quad (24)$$

We may picture this as a hypersurface of real dimension  $\leq n$  embedded in the unit sphere in  $\mathcal{H}$ ; evidently  $G$  acts transitively on  $\vartheta(\psi_0)$ . Next we define two stability groups  $H_0, H$  associated with  $\psi_0$ :

$$H_0 = \{g \in G | \mathcal{U}(g)\psi_0 = \psi_0\} \subset G; \quad (25)$$

$$H = \{g \in G | \mathcal{U}(g)\psi_0 = (\text{phase factor})\psi_0\} \subset G \quad (26)$$

$H_0$  is the stability group of  $\psi_0$  in the strict sense, while  $H$  is the stability group of  $\psi_0$  upto phases. Both are subgroups of  $G$ , and moreover  $H_0$  is a normal subgroup of  $H$ . We can see that the relationship between  $H_0$  and  $H$  must be one of three possibilities, depending on the nature of the factor group  $H/H_0$ :

$$\begin{aligned} (i) H/H_0 &= \text{trivial;} \\ (ii) H/H_0 &= \text{nontrivial discrete;} \\ (iii) H/H_0 &\simeq U(1). \end{aligned} \quad (27)$$

Turning to the Lie algebras, in cases (6.4 (i), (ii)) it is clear that  $H_0$  and  $H$  share the same generators. Let us denote them by  $T_a$ , where  $a = 1, 2, \dots, k = \dim H_0$ ; so these obey a set of commutation relations on their own, involving the structure constants of  $H_0$ , and they all annihilate  $\psi_0$ :

$$[T_a, T_b] = i f_{ab}{}^c T_c, \quad a, b, c = 1, \dots, k; \quad (28)$$

$$T_a \psi_0 = 0, \quad a = 1, 2, \dots, k \quad (29)$$

At this point we recognize that the (real) dimension of  $\vartheta(\psi_0)$  is  $(n - k)$ . We can now add further generators  $T_\mu$  of  $G$ , where  $\mu$  takes on  $(n - k)$  distinct values. One then has the additional information:

$$\underline{G} = \text{span } \{T_a, T_\mu\},$$

$$C^\mu T_\mu \psi_0 = \psi_0 \iff C^\mu = 0 \quad (30)$$

At this point we make an important assumption which is indeed valid in all physically important situations. This is the assumption that  $H_0$  is a reductive subgroup in  $G$ . This has the consequence that the additional elements  $T_\mu$  needed to make up a basis for the Lie algebra  $\underline{G}$  of  $G$  can be chosen so that we have the commutation relations

$$[T_a, T_\mu] = i f_{a\mu}{}^\nu T_\nu, \quad (31)$$

the point being that no generators  $T_b$  of  $H_0$  appear on the right hand side. In finite terms this is the statement that the  $T_\mu$  furnish some collection of irreducible tensor operators with respect to  $H_0$ . This conclusion is also obtained in case  $H_0$  is compact, as this guarantees the full reducibility of any representation of  $H_0$ ; this is a special case of being reductive.

In the case (6.4(iii)) the situation is a bit different. The generators of  $H$  consist of the generators  $T_a$  of  $H_0$ , and one extra generator which we denote as  $Y$  and which can be assumed to be normalised in a particular way. The relevant formulae in this case are:

$$\begin{aligned} T_a \psi_0 &= 0, \quad Y \psi_0 = \psi_0; \\ [T_a, Y] &= 0 \end{aligned} \quad (32)$$



Thus  $Y$  may be counted as one of the  $T_\mu$ , which happens to be an  $H_0$ -scalar generator.

One more concept is needed at the Lie algebra level before we can turn to dynamical phases. We denote by  $\{T_\rho\}$  all those generators of  $G$  which are scalars with respect to the stability group  $H$ , and which more over lie outside of  $H_0$ . Thus in terms of the three possibilities listed in eqn.(6.4) for the relationship between  $H_0$  and  $H$ , we have:

Cases (6.4(i),(ii)) :  $T_\rho = T_\mu$  such that

$$[T_a, T_\rho] = 0; \quad (33)$$

Case (6.4 (iii)):  $T_\rho = T_\mu$  such that

$$\begin{aligned} [T_a, T_\rho] &= [Y, T_\rho] = 0, \\ Y &= \text{one of the } T_\rho. \end{aligned} \quad (34)$$

The importance of these generators of  $G$  is that, because of the Wigner-Eckart theorem and keeping in mind eqn.(6.8),

$$(\psi_0, T_r \psi_0) \neq 0 \implies T_r = \text{one of the } T_\rho \quad (35)$$

Now let us specify the Hilbert space curves  $\mathcal{C}$  for which we wish to examine the dynamical phase  $\varphi_{\text{dyn}}[\mathcal{C}]$ . These are curves lying in the orbit  $\mathcal{O}(\psi_0)$ , produced by a parametrised curve  $\{g(s)\}$  of group elements acting on  $\psi_0$ :

$$\mathcal{C} = \{\psi(s) \equiv \psi(g(s)) = \mathcal{U}(g(s))\psi_0 | g(s) \in G, g(0) \in H_0, 0 \leq s \leq s_1\}. \quad (36)$$

Here we have let the parameter  $s$  run from 0 to some  $s_1$ , and have ensured that the curve starts out from  $\psi_0$  by demanding  $g(0)$  be in the subgroup  $H_0$ . Now by standard and wellknown arguments we know that the orbit  $\mathcal{O}(\psi_0)$  can be identified in a natural way with the coset space  $G/H_0$ . This gives us the freedom to either think of the curve  $\mathcal{C}$  as embedded in Hilbert space, or as lying in the coset space  $G/H_0$ . The former view makes available the unitary representation  $\mathcal{U}(g)$  of  $G$  and its generators; while the latter makes available the rich differential geometric structures known to exist on  $G$  and on  $G/H_0$ . Correspondingly it now turns out that there is a clean separation of  $\varphi_{\text{dyn}}[\mathcal{C}]$  into algebraic and geometric ingredients. On the manifold of the group itself we know that there are left and right invariant vector fields,  $X_r^{(0)}$  and  $\widehat{X}_r^{(0)}$  say, generating the action of  $G$  on itself by right and by left translations respectively. Dual to these sets of vector fields are the two sets of one forms, the Maurer-Cartan one forms, conventionally written as  $\theta^{(0)r}$  and  $\widehat{\theta}^{(0)r}$  and enjoying respectively left and right translation invariance. These forms can be viewed as paired in a reciprocal manner with the generators  $T_r$  of the representation  $\mathcal{U}(\cdot)$  of  $G$ , in fact with a basis for the Lie algebra  $\underline{G}$ . The Lie algebra commutation relations among the vector fields are mirrored in the Maurer-Cartan relations obeyed by  $\theta^{(0)r}$  and  $\widehat{\theta}^{(0)r}$ . Now when we descend from the group  $G$  to the coset space  $G/H_0$ , (and for definiteness we take  $G/H_0$  to be the space of right cosets  $gH$ ), the right invariant vector fields generating left translations do project down properly to globally defined vector fields  $X_r$  on  $G/H_0$ , generating the transitive action of  $G$  on  $G/H_0$ . The one forms  $\theta^{(0)r}, \widehat{\theta}^{(0)r}$  in general give rise to only locally defined one forms over  $G/H_0$ , except for the  $\widehat{\theta}^{(0)\rho}$  which are paired with (or go with) the  $H$ -scalar generators  $T_\rho$  lying outside  $H_0$ . These lead to a set of one-forms  $\widehat{\theta}^\rho$  on  $G/H_0$  which are globally well-defined.

These one-forms  $\widehat{\theta}^\rho$  are crucial for dynamical phase calculations. Indeed one finds, after a careful calculation, the interesting and compact result

$$\varphi_{\text{dyn}}[\mathcal{C}] = -(\psi_0, T_\rho \psi_0) \int_{\mathcal{C} \subset G/H_0} \widehat{\theta}^\rho. \quad (37)$$

Here a sum on the repeated index  $\rho$  is implied. As indicated earlier, the standard Wigner-Eckart theorem of quantum mechanics plays a role in the derivation of this result, and in this context the

property of the generators  $T_\rho$  expressed by eqn.(6.10) becomes relevant. Beyond this practical use of the Wigner-Eckart Theorem, there is also a similarity in spirit between that theorem and the final result (6.12) for dynamical phases, in the sense that there is a clean separation into algebraic representation dependent factors, and differential geometric  $\mathcal{C}$ -dependent factors. Indeed, even the dependence on the fiducial vector  $\psi_0$  occurs only in the former factors.

The result (6.12) is of course one ingredient in the calculation of geometric phases, the other being the Pancharatnam phase term. In any event, for a given group  $G$ , if one has a list of all possible subgroups  $H_0$  upto conjugation, the result (6.12) enables one to quickly determine those situations wherein nontrivial dynamical phases can occur; if due to the selection rules implicit in (6.12) we have a vanishing dynamical phase, we have a trivial situation wherein the geometric phase reduces to the nonlocal Pancharatnam contribution.

Applications of the result (6.12) to the groups  $SU(3)$ ,  $Sp(2)$  and even to the familiar cases of  $SU(2)$  and  $SO(3)$ , may be found in the literature[18].

## 7 Concluding remarks

We have given a brief account of the theory of the geometric phase emphasizing two aspects - the organic links with the much earlier work of Pancharatnam and of Bargmann, and the deep connection to group actions and Lie group representations. The perceptive reader would not fail to be impressed by the subtlety and richness of the connections between the ideas of Pancharatnam, Bargmann and Berry; and it is certainly very satisfying to see how the final form of the geometric phase concept has brought all these strands together. We mentioned that Bargmann's work was in the context of the Wigner unitary-antiunitary theorem for the representation of symmetry operations in quantum mechanics. It is but fair to say here that recently a new approach to the Wigner theorem based on Pancharatnam's ideas has been presented [19].

The geometric phase concept has been exploited in the quantum field theoretic context where it links up to the so-called anomalies. A great deal of work has also been done to unravel the differential geometric features of this phase[20]. To keep this article within bounds, as well as to focus on the two main aspects mentioned above, we have limited ourselves to an account within the framework of basic quantum mechanics.

## References

- [1] M.V. Berry, Proc. Roy. Soc. A 392, 45(1984).
- [2] A very useful reprint collection on the entire subject is "Geometric Phases in Physics", A Shapere and F. Wilczek (eds.), World Scientific Publishing Co., Singapore (1989).
- [3] M.V.Berry, "Quantum adiabatic anholonomy", Lectures given at the Ferrara School of Theoretical Physics on "Anomalies, defects, phases....", June 1989.
- [4] S.Pancharatnam, Proc. Indian Acad. Sci A 44, 247 (1956); attention was called to this important work by S. Ramaseshan and R. Nityananda, Curr. Sci. 55, 1225 (1986).
- [5] V. Bargmann, Jour. Math. Phys. 5, 862 (1964).
- [6] Y.Aharonov and J. Anandan, Phys. Rev. Lett. 58, 1593 (1987).
- [7] J. Samuel and R. Bhandari, Phys. Rev. Lett. 60, 2339 (1988).
- [8] N.Mukunda and R. Simon, Ann. Phys. (NY) 228, 205 (1993).
- [9] N.Mukunda and R. Simon, Ann. Phys. (NY) 228, 269 (1993).
- [10] R.Y. Chiao and Y.S. Wu, Phys. Rev. Lett. 57, 933 (1986).

- [11] R. Bhandari and J. Samuel, Phys. Rev. Lett. 60, 1210 (1988).
- [12] R. Simon, H.J. Kimble and E.C.G. Sudarshan, Phys. Rev. Lett. 61, 19 (1988).
- [13] E.M.Rabei, Arvind, N.Mukunda and R.Simon, "Bargmann Invariants and Geometric Phases - A Generalised Connection", Phys. Rev. A (in press).
- [14] S. Chaturvedi, M.S. Sriram and V. Srinivasan, J. Phys. A 20, L1071 (1987).
- [15] R.Simon and N.Mukunda, Phys. Rev. Lett. 70, 880 (1993).
- [16] For a very lucid account of Pancharatnam's own ideas, see R. Nityananda, Current Science 67, 238 (1994).
- [17] G.Khanna, S.Mukhopadhyay, R.Simon and N.Mukunda, Ann. Phys (NY) 253, 55 (1997); Arvind, K.S. Mallesh and N. Mukunda, J. Phys. A 30, 2417 (1997).
- [18] See, for instance, N.Mukunda "Group Theoretical Aspects of the Geometric Phase in Quantum Mechanics", invited plenary lecture at the International Conference on Group Theoretical Methods in Physics, ICGTMP-98, University of Tasmania, Hobart, Tasmania, Australia, July 1998.
- [19] J.Samuel, Pramana - Journal of Physics 48, 959 (1997).
- [20] The earliest step in this direction, B. Simon, Phys. Rev. Lett. 51, 2167 (1983), actually appeared in print before reference (1)!

# 13. The Skyrme Model for Baryons

J. Schechter<sup>a</sup> and H. Weigel<sup>†, b \*</sup>

<sup>a)</sup>Department of Physics, Syracuse University  
Syracuse, NY 13244-1130

<sup>b)</sup>Center for Theoretical Physics  
Laboratory of Nuclear Science and Department of Physics  
Massachusetts Institute of Technology  
Cambridge, Ma 02139

## Abstract

We review the Skyrme model approach which treats baryons as solitons of an effective meson theory. We start out with a historical introduction and a concise discussion of the original two flavor Skyrme model and its interpretation. Then we develop the theme, motivated by the large  $N_C$  approximation of QCD, that the *effective* Lagrangian of QCD is in fact one which contains just mesons of all spins. When this Lagrangian is (at least approximately) determined from the meson sector it should then yield a zero parameter description of the baryons. We next discuss the concept of chiral symmetry and the technology involved in handling the three flavor extension of the model at the collective level. This material is used to discuss properties of the light baryons based on three flavor meson Lagrangians containing just pseudoscalars and also pseudoscalars plus vectors. The improvements obtained by including vectors are exemplified in the treatment of the *proton spin puzzle*.

\*This work is supported in parts by funds provided by the U.S. Department of Energy (D.O.E.) under cooperative research agreements #DR-FG-02-92ER420231 & #DF-FC02-94ER40818 and the Deutsche Forschungsgemeinschaft (DFG) under contracts We 1254/3-1 & 1254/4-1.

<sup>†</sup>Heisenberg-Fellow

## 1 Historical background and motivation

The Skyrme model was born around 1960 in a series of increasingly more detailed papers [1]. At that time the prevailing dynamical model of nuclear forces was that of Yukawa which had been formulated in the 1930's. Still to come was the concept of fractionally charged quarks and much further in the future was the recognition that the correct theory of strong interactions binds these quarks together with non-Abelian (color) gauge fields.

In the Yukawa theory, of course, the nucleons are introduced as fundamental fermion fields while the spin zero pion fields are postulated to provide the "glue" which binds protons and neutrons into nuclei. This model is acknowledged to work reasonably well as a description of the long range interactions of nucleons and the prediction of the existence of pions has been amply confirmed.

Skyrme's innovation was to provide a model in which the fundamental fields consisted of just the pions. The nucleon was then obtained, in the initial approximation, as a certain classical configuration of the pion fields. The seeming contradiction of making fermi fields out of bose fields was avoided by arranging the classical field configuration to possess a non-zero "winding number". In modern language this "Skyrmion" is an example of a topological soliton. Such objects are

solutions to the classical field equations with localized energy density [2]. They play an important role nowadays in many areas of physics and the papers of Skyrme are justifiably recognized as pioneering milestones in this development.

The years following this original idea saw the particle physics community actively investigating the approaches of quark models, flavor symmetry, current algebra, chiral dynamics, dual resonance models and finally color gauge theory to the problem of strong dynamics. Evidently Skyrme's model was lost in the rush. However the novelty of the model did stimulate a few interesting papers [3, 4, 5, 6] before the more recent wave of activity in the area.

At first glance, it might appear that a Lagrangian model built out of only pion fields could not be more different as a description of the nucleons from the current picture of three "valence" quarks containing a trivalent "color" index and bound together through their interaction with  $SU(3)$  gauge fields. Remarkably, it has turned out that the Skyrme model is in fact a plausible approximation to this QCD picture. This may be understood as follows.

In QCD the gauge coupling constant has an effective strength which decreases for interactions at high energy scales (asymptotic freedom) but which increases at the low energy scales which are relevant when one considers the binding of quarks into nucleons and other hadrons. Thus the application of standard perturbation theory techniques to the problem of low energy interactions is not expected to be reliable and in fact has not produced definitive results. A natural alternative approach which retains the possibility of using perturbation theory is to imagine that the strong underlying gauge couplings bind the quarks into particles which may possibly interact with each other relatively weakly. At low energies these particles should evidently comprise the pseudoscalar meson fields (pions when restricted to two "flavors"). Then it is necessary to formulate some *effective* Lagrangian for the pions. Certainly the Lagrangian should be restricted by the correct symmetries of the underlying gauge theory. These must include an (approximate)  $SU(N_f)$  flavor symmetry [7], where  $N_f$  is the number of light flavors.

But there is another symmetry which plays a crucial role. At about the same time that Skyrme was contemplating the model under discussion the correct formulation [8] of the structure of the effective weak (beta-decay etc.) interaction was discovered. It was noted that this interaction treated the left and right handed components of fermions on a completely separate basis. If this distinction is maintained at the level of the strong interactions one should impose a "chiral" left handed  $SU(3)$  flavor  $\times$  right handed  $SU(3)$  flavor symmetry on the effective low energy Lagrangian of mesons. A consequence of this larger symmetry is that the meson multiplets must contain scalar as well as the desired low-lying pseudoscalar particles. This seemed a bit of an embarrassment in that the pseudoscalars are very light while the scalars were not well established and presumably heavy. Hence the possibility of a degenerate symmetry multiplet is implausible. Nevertheless it was realized [9] that the situation was likely to be similar to that met in the BCS theory of superconductivity in which the vacuum (ground state) is energetically favored to exist in a non-symmetric state. This "spontaneous breakdown" picture predicts, in the absence of a needed small explicit symmetry breaker, exactly zero mass for the pseudoscalars at the same time that the scalars are massive. In fact it may be formulated using a so called "non-linear realization of chiral symmetry" in such a way that the scalars do not appear at all [10]. The prototype Lagrangian density for this picture is

$$\mathcal{L} = \frac{f_\pi^2}{4} \text{tr} (\partial_\mu U \partial^\mu U^\dagger) + \dots, \quad (1)$$

where  $U = \exp(\sqrt{2}i\phi/f_\pi)$ ,  $\phi$  being the  $3 \times 3$  matrix of the ordinary pseudoscalar mesons and  $f_\pi = 93\text{MeV}$  the "pion decay constant".  $U$  is a unitary matrix which transforms "linearly" under the chiral transformations. Possible higher derivative and symmetry breaking terms have not been explicitly written here. It was demonstrated a long time ago that just this term compactly summarizes the low energy scattering of pseudoscalar mesons [11]. Improvements to this term form the basis of the "chiral perturbation theory approach" [12]. Now it is believed that a picture like this is expected from fundamental QCD. However the same Lagrangian was earlier written by Skyrme (in the two rather than three flavor case) in order to explain the nucleons [1] before the present justifications for it were known.

Even in the framework of the chiral Lagrangian given above it would seem that there is no special *a priori* reason not to explicitly add baryons in a chiral symmetric manner rather than to build them out of the mesons. Indeed there have been many papers over very many years which do just this with reasonable phenomenological results [13]. Nevertheless there is an indication from fundamental QCD that the soliton treatment of the baryon is more natural. This arises from an attempt [14] to consider  $1/N_C$ , the inverse of the number of colors in the gauge theory, as a possible expansion parameter for QCD which might be meaningful even at low energies. In this approach the product  $g'^2 = g^2 N_C$ , where  $g$  is the gauge theory coupling constant, is held constant. 't Hooft [14] showed that for large  $N_C$  QCD may be considered as a theory of mesons weakly interacting in the sense that scattering amplitudes or quadrilinear coupling constants are of order  $g_{\text{eff}} = 1/N_C$ . Since the baryon mass must start out proportional to  $N_C$  (noting that the baryon in the  $N_C$  model is made of  $N_C$  quarks) it means that the predicted expressions for baryon masses should start out as the inverse of this coupling constant. In the framework of the (non-relativistic) mean field treatment Witten [15] not only pointed out that the baryon masses indeed grow linearly with  $N_C$  but also that baryon radii and meson-baryon scattering amplitudes are of the order  $N_C^0$  while baryon-baryon scattering is of  $\mathcal{O}(N_C)$ . He in particular recognized that this inverse behavior with  $g_{\text{eff}}$  is just the usual signal that the baryon state in question is a soliton of the effective meson theory.

Naturally, one wonders how these “modern” justifications for the Skyrme approach relate to Skyrme’s original motivations. We are fortunate in having available a reconstructed talk on just this topic by Skyrme [16]. He mentioned three motivations: 1) The idea of unifying bosons and fermions in a common framework. 2) The feeling that point particles are inconsistent in the sense that their quantum field theory formulation introduces infinities which are only “swept under the rug” by the renormalization process. 3) The desire to eliminate fermions from a fundamental formulation since fermions have no simple classical analog. What seems more fascinating is his awareness that there were probably some “hidden” influences pushing him toward the soliton picture. Directly, these came from his fascination with Kelvin’s idea that the various atoms should correspond to vortices of different connectivities in some underlying liquid. In turn, his interest in Kelvin was sparked at an early age by the presence of a tide prediction machine, designed by Kelvin and constructed by his great-grandfather, still occupying space in his great-grandfather’s house. An interesting account of this aspect is given in a paper of Dalitz [17].

Thus it seems that Skyrme’s motivations were not those currently used to justify his model. In particular it appears that he did not choose his Lagrangian model to describe spontaneous breakdown of chiral symmetry. Rather the non-linear form was chosen to insure that the pions were “angular” variables which would give multi-valued functions; the crossing of different sheets of these functions might then correspond to singularities which would realize the baryons. The evident “moral” of this historical discussion is just that interesting ideas have an uncanny way of turning out to be useful and true. In this spirit, we would like to continue with the application of Skyrme’s ideas to current research on baryon physics, making use of current motivations but trying to avoid getting enslaved by them.

## 2 The Skyrme model for two flavors

In this section we will present the basic technology of the Skyrme model for baryons. The starting point for the construction of a soliton solution is the non-linear sigma model Lagrangian (1) already introduced in the previous section. As we require a finite energy density the chiral field  $U$  must approach a constant value at spatial infinity. We are free to choose this to be unity, *i.e.*,

$$U(\mathbf{r}, t) \xrightarrow{|\mathbf{r}| \rightarrow \infty} 1. \quad (2)$$

This can be considered a mapping from compactified coordinate space, a three-sphere  $S^3$ , to the space which is described by the unitary, unimodular matrix  $U$ , namely  $SU(N_f)$ , where  $N_f$  denotes the number of flavors. In the case of two flavors the target space is isomorphic to  $S^3$ . The mappings  $S^3 \rightarrow S^3$  fall into distinct equivalence classes. This signals the existence of soliton configurations

because members of different classes cannot be continuously transformed into one-another. The equivalence classes are characterized by the winding number. This number counts the coverings of the target space and is the charge  $\int d^3x B_0$  associated with the topological current

$$B_\mu = \frac{1}{24\pi^2} \epsilon_{\mu\nu\rho\sigma} \text{tr} [(U^\dagger \partial^\nu U) (U^\dagger \partial^\rho U) (U^\dagger \partial^\sigma U)] . \quad (3)$$

When later discussing the three flavor case we will see that this topological current indeed equals the baryon number current.

Although these topological considerations allow the existence of soliton solutions it turns out that the dynamics of (1) do not lead to static stable classical solutions. This can be deduced from a simple consideration, known as Derrick's theorem [18]. Assume  $U_0(\mathbf{r})$  to be such a solution. The static energy of  $U_0(\lambda\mathbf{r})$ , obtained from the Hamiltonian of (1.1), would then be

$$E_{\text{cl}}^{(\text{nl}\sigma)}[U_0(\lambda\mathbf{r})] = \frac{1}{\lambda} E_{\text{cl}}^{(\text{nl}\sigma)}[U_0(\mathbf{r})] \quad (4)$$

which does not have a minimum at  $\lambda = 1$ , in contradiction to the assumption. In order to obtain stable solitons Skyrme added a term to the Lagrangian which is of fourth order in the derivatives,

$$\mathcal{L}^{(\text{Sk})} = \frac{1}{32e^2} \text{tr} ([U^\dagger \partial_\mu U, U^\dagger \partial_\nu U] [U^\dagger \partial^\mu U, U^\dagger \partial^\nu U]) . \quad (5)$$

Here  $e$  is the dimensionless "Skyrme constant". Although this term is quartic in the derivatives it was arranged to be at most quadratic in the time-derivatives. This makes the quantization feasible. It is now apparent that a scaled configuration may well lead to a minimum of the energy functional

$$E_{\text{cl}}^{(\text{tot})}[U_0(\lambda\mathbf{r})] = \frac{1}{\lambda} E_{\text{cl}}^{(\text{nl}\sigma)}[U_0(\mathbf{r})] + \lambda E_{\text{cl}}^{(\text{Sk})}[U_0(\mathbf{r})] \quad (6)$$

at  $\lambda = 1$  provided the configuration  $U_0(\mathbf{r})$  satisfies  $E_{\text{cl}}^{(\text{nl}\sigma)}[U_0(\mathbf{r})] = E_{\text{cl}}^{(\text{Sk})}[U_0(\mathbf{r})]$ .

*A priori* the Euler-Lagrange equations of motion for the chiral field  $U_0(\mathbf{r})$  are highly non-linear partial differential equations. To simplify these equations Skyrme adopted the famous hedgehog *ansatz*

$$U_0(\mathbf{r}) = \exp(i\boldsymbol{\tau} \cdot \hat{\mathbf{r}} F(r)) , \quad (7)$$

where  $\boldsymbol{\tau}$  represents the Pauli matrices. This form may actually be traced back to the old "strong coupling" theory [19]. Upon substitution of this *ansatz* the energy functional turns into a simple integral involving only the radial function  $F(r)$ ,

$$E[F] = \frac{2\pi f_\pi}{e} \int_0^\infty dx \left\{ (x^2 F'^2 + 2\sin^2 F) + \sin^2 F \left( 2F'^2 + \frac{\sin^2 F}{x^2} \right) \right\} . \quad (8)$$

Henceforth this radial function will be called the chiral angle. In eq (8) a prime indicates a derivative with respect to the dimensionless coordinate  $x = ef_\pi r$ . In this manner we have completely extracted the dependence on the model parameters. Imposing  $F(\infty)=0$  and noting that  $\int d^3r B_0 = (F(0) - F(\infty))/\pi$  leads to the boundary condition  $F(0) = \pi$  for a unit baryon number configuration. The profile function depicted in Fig. 2.1 minimizes (8) and is obtained numerically. The energy obtained by substituting this solution into (8) is found to be  $E = 23.2\pi f_\pi/e$  [6].

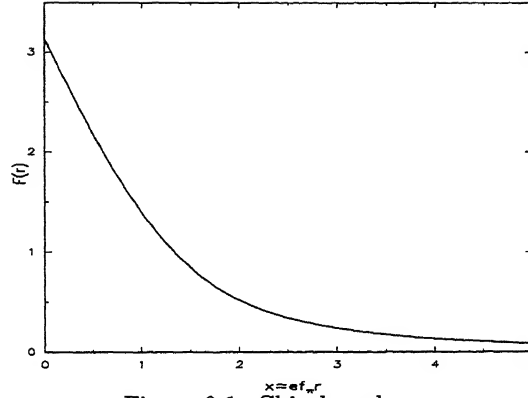


Figure 2.1: Chiral angle

As the *ansatz* (7) is not invariant under separate spatial or flavor rotations this field configuration does not yet describe states with good spin and flavor quantum numbers. As the first step towards generating such states, time-dependent collective coordinates  $A(t)$  are introduced which describe the spin and flavor orientation of the hedgehog,

$$U(\mathbf{r}, t) = A(t)U_0(\mathbf{r})A^\dagger(t), \quad A(t) \in SU(N_f). \quad (9)$$

Note that the hedgehog structure causes rotations in coordinate and flavor space to be equivalent. For generality we assumed an arbitrary number of flavors. In the two flavor case this configuration yields the Lagrange function

$$L(A, \dot{A}) = \frac{1}{2}\alpha^2\Omega^2 - E_{\text{cl}} \quad (10)$$

where the quantity  $\Omega$  measures the time dependence of the collective coordinates

$$A^\dagger(t)\frac{d}{dt}A(t) = \frac{i}{2}\boldsymbol{\tau} \cdot \boldsymbol{\Omega}. \quad (11)$$

The constant of proportionality,  $\alpha^2$  is computed as a spatial integral over the chiral angle to be  $\alpha^2 = 53.3/(e^3 f_\pi)$ . Computing the spin as the Noether charge of spatial rotations yields  $\mathbf{J} = \alpha^2\boldsymbol{\Omega}$ . Apparently the system displays all the features of a rigid top. In that language  $\boldsymbol{\Omega}$  and  $\alpha^2$  are denoted as the angular velocity and the moment of inertia, respectively.

In the second step the collective coordinates are elevated to quantum variables. Again this is completely analogous to the quantization of the rigid top and gives the quantization rule  $[J_i, J_j] = i\epsilon_{ijk}J_k$ . For the hedgehog *ansatz* in  $SU(2)$  spin and isospin are related via the adjoint representation of the collective coordinates, i.e.  $I_i = -D_{ij}J_j$  with  $D_{ij} = (1/2)\text{tr}(\tau_i A \tau_j A^\dagger)$  due to the equivalence of the respective rotations. Hence only states which have identical spin and isospin are allowed in the spectrum. These are the nucleon ( $I = J = 1/2$ ) and the  $\Delta$ -resonance ( $I = J = 3/2$ ). Finally the Hamiltonian for the collective coordinates is given by

$$H_{\text{coll}} = E_{\text{cl}} + \frac{1}{2\alpha^2}\mathbf{J}^2 = E_{\text{cl}} + \frac{1}{2\alpha^2}\mathbf{I}^2 \quad (12)$$

which yields the  $\Delta$ -nucleon mass difference

$$M_\Delta - M_N = \frac{3}{2\alpha^2}. \quad (13)$$

Using the physical value for the pion decay constant,  $f_\pi = 93\text{MeV}$  requires us to choose  $e \approx 4.75$  to reproduce the empirical mass difference of  $293\text{MeV}$ . Substituting  $e \approx 4.75$  into eq (8) yields the classical nucleon energy  $E = 23.2\pi f_\pi/e \approx 1430\text{MeV}$ . This is not in especially good agreement with the experimental value of about  $939\text{MeV}$ . However the following points must be kept in mind:



- (i) The meson Lagrangian consisting of eq (1) plus eq (5) contains only pseudoscalars. We would expect that other low mass mesons (notably the vectors) should also be included. The large  $N_C$  expansion [14, 15] requires an infinite number but common sense suggests a reasonable approximation for explaining hadronic physics up to about 1GeV would be to keep those mesons with masses up to this value. Certainly the predictions in the mesonic sector of the theory are noticeably improved by the inclusion of vector mesons. The consistency of the overall picture requires accurate predictions both in the mesonic and baryonic sectors of the effective theory.
- (ii) In nature there are three rather than two “light” flavors and this aspect should be included in a realistic formulation. (This feature also makes more transparent the origin of the topological current eq (3).) Furthermore the effects of flavor and chiral symmetry breaking mediated by the finite values of the quark masses have not yet been taken into account.
- (iii) Order of  $N_C^0$  corrections to the nucleon mass which have the structure of the Casimir effect in field theory have also not been included. These quantum contributions to the energy have been estimated to be negative and of the order of a few hundred MeV, predicting a total nucleon mass at the order of the experimental value [20]. Nevertheless one should be cautious about these quantum corrections, after all the Skyrme model is not renormalizable, leaving a logarithmic scale dependence of the “renormalized” Casimir energy. It seems that at best the quantum corrections can be computed in a scenario compatible with the chiral expansion.

We will postpone the discussion of a variety of nucleon (and other baryons’) properties until after we have treated the more general case of flavor  $SU(3)$ .

To end this section on the basics of the Skyrme model we would like to briefly discuss the consistency of the Skyrme model with the large  $N_C$  picture of QCD. In section 1 we have already noted that the quadrilinear coupling between mesons scales like  $1/N_C$ . To check this behavior it is convenient to expand the non-linear  $\sigma$  model Lagrangian (1) in powers of the pion field:

$$\frac{1}{2}\partial_\mu\pi\cdot\partial^\mu\pi + \frac{1}{6f_\pi^2}\left\{(\pi\cdot\partial_\mu\pi)^2 - \pi^2\partial_\mu\pi\cdot\partial^\mu\pi\right\} + \mathcal{O}(\pi^6) . \quad (14)$$

Since the quadrilinear coupling constant is  $1/f_\pi^2$  we deduce that  $f_\pi \sim \sqrt{N_C}$ . This agrees with general arguments [15]. Similarly the Skyrme term (5) provides a quartic pion interaction with the coupling constant  $1/(e^2 f_\pi^4)$  which implies  $e \sim 1/\sqrt{N_C}$ . Hence the classical energy (8) grows linearly with the number of colors as asserted from the corresponding generalization of QCD. Moreover, without flavor symmetry breaking large  $N_C$  QCD predicts the baryons of different  $J$  to be degenerate [21]. This is perfectly consistent with the mass formula (12) because the moment of inertia also grows linearly with  $N_C$  as is indicated after eq (11), hence the second term in (12) behaves like  $1/N_C$ .

Actually, the understanding of the  $N_C$  expansion for baryons involves some subtleties. Consider the construction of large  $N_C$  baryons in the quark model. The lowest lying baryons are made of  $N_C$  (taken to be odd) quarks in a totally antisymmetric (*i.e.* singlet) color spin state with no orbital angular momentum. One expects particles of all total angular momenta from  $J = 1/2$  to  $J = N_C/2$  to be obtained. In agreement with the spectrum of eq (12) we expect  $I = J$  for these particles and an infinite number of them as  $N_C \rightarrow \infty$ . The trouble is that there is no experimental evidence for any  $I = J = 5/2, 7/2$  etc. particles. This may be interpreted as evidence that  $N_C = 3$  in nature. Still, the large  $N_C$  expansion is useful if one computes a quantity which exists in the  $N_C = 3$  theory as a (presumably quickly convergent) Taylor series in  $1/N_C$ . For the specific case of the higher excitations  $J = 5/2, 7/2, \dots$  the above treatment of the rotational modes seems inadequate because the rotational energy gets as large as the classical contribution. By including these modes in the Euler–Lagrange equations the widths of these higher excitations have been estimated to be comparable to their masses [22]. This makes a particle interpretation of these states problematic suggesting that they are artifacts of the collective quantization method employed rather than of physical relevance. Possible caveats for these calculations are the instability of these configurations against emitting pions [23] and that the results are only obtained by analytical continuation in the spin variable.

### 3 Chiral symmetry and its breaking

In this section we will briefly discuss the concept of chiral symmetry which represents a guiding principle for extending the Skyrme model. Attention will be limited to those aspects of this large subject which have direct relevance to the study of Skyrmions. The basic idea is to construct a model of meson fields which “mocks up” as many symmetries and properties of the fundamental QCD Lagrangian as possible.

#### 3.1 The QCD Lagrangian

Let us first recall the matter piece of the QCD Lagrangian

$$\begin{aligned}\mathcal{L}_{\text{QCD}}^{\text{matter}} &= \sum_{f=1}^{N_f} \bar{q}_f (i\partial + g\mathcal{A} - m_f) q_f \\ &= \sum_{f=1}^{N_f} \{ \bar{q}_{f,L} (i\partial + g\mathcal{A}) q_{f,L} + \bar{q}_{f,R} (i\partial + g\mathcal{A}) q_{f,R} - m_f (\bar{q}_{f,L} q_{f,R} + \bar{q}_{f,R} q_{f,L}) \} .\end{aligned}\quad (15)$$

Here  $\mathcal{A}_\mu$  is the matrix representation of the gluon fields and  $g$  the quark–gluon coupling. Most notably we have introduced the chiral representation for the QCD current quarks

$$q_{L,R} = \frac{1}{2} (1 \mp \gamma_5) q \quad (16)$$

of each flavor.

Strictly speaking, the quark mass terms are not part of the QCD Lagrangian but arise from the Yukawa terms of the full microscopic theory of nature. A major unsolved problem is to understand the resulting pattern of quark mass parameters. The phenomenologically determined masses [24] are  $m_u \approx 5\text{MeV}$ ,  $m_d \approx 9\text{MeV}$ ,  $m_s \approx 120 - 170\text{MeV}$ ,  $m_c \approx 1.5\text{GeV}$ ,  $m_b \approx 4.5\text{GeV}$  and  $m_t \approx 175\text{GeV}$ . This random looking perturbation of the “strong interaction” plays a crucial role in determining the nature of elementary particle physics. In the region up to about  $1\text{GeV}$  it is not possible to produce particles containing  $c$ ,  $b$  or  $t$  quarks. Then it is usually a good approximation to simply drop them from the theory. Other approximations are useful when dealing with the subspace carrying the flavor quantum number of a single “heavy” quark [25]. Furthermore in the sector of the three “light” quarks  $u, d, s$  it turns out to be fundamental to neglect the  $u, d, s$  masses as a first approximation and include their effects as a perturbation [26]. This is reasonable because the light masses are less than the quantity  $\Lambda_{\text{QCD}} \approx 250\text{MeV}$ , the scale below which the QCD effective coupling gets extremely large.

In the case  $m_f = 0$  the Lagrangian (15) specialized to the three light quarks has the global chiral symmetry

$$U_L(3) \times U_R(3) : \quad q_L \longrightarrow L q_L \quad \text{and} \quad q_R \longrightarrow R q_R \quad \text{with} \quad q_{L,R} = \begin{pmatrix} q_u \\ q_d \\ q_s \end{pmatrix}_{L,R} \quad (17)$$

where  $L$  and  $R$  are each  $3 \times 3$  unitary matrices. Using Noether’s theorem on the classical Lagrangian then yields the conservation of the eighteen vector and axial vector currents

$$j_{ij}^\mu = \bar{q}_j \gamma^\mu q_i \quad \text{and} \quad j_{ij,5}^\mu = \bar{q}_j \gamma^\mu \gamma_5 q_i, \quad (18)$$

where the latin indices run over  $u, d, s$ . These currents play an important role in the theory of weak interactions.

Now a major discovery of quantum field theory is that consequences of the classical field equations of motion (which can be used to verify the conservation of the Noether currents) do not necessarily hold at the quantum level. It is necessary to consider whether there exists a suitable regularization of the divergent diagrams of the theory which maintains the classical relations. In the present case, the axial singlet current<sup>1</sup>,  $j_\mu^5 = \bar{q} \gamma_\mu \gamma_5 (\lambda^0/2) q$  is not conserved even for massless

<sup>1</sup>Here  $\lambda^a$ ,  $a = 1, \dots, N_f^2 - 1$  denote the Gell–Mann matrices of  $SU(N_f)$  while  $\lambda^0$  refers to the singlet generator.

quarks. Rather its divergence is proportional to the gluon field tensor times its dual. This is a result of the well-known Adler–Bell–Jackiw (ABJ) triangle anomaly [27] contained in the loop–diagrams shown in figure 1.

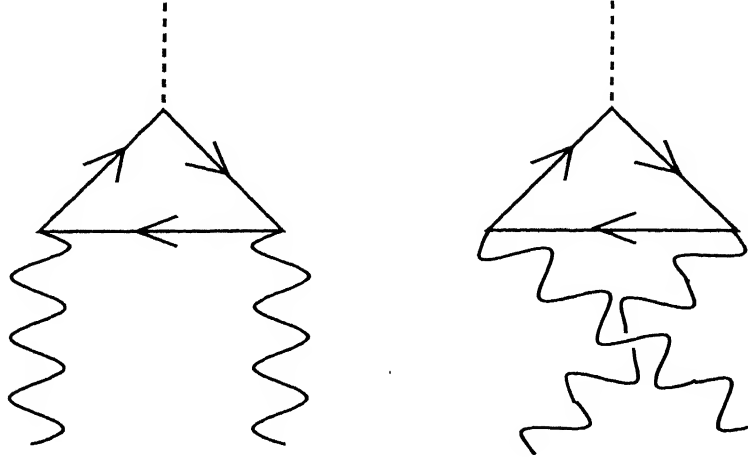


Figure 1: Adler–Bell–Jackiw anomaly. The arrows indicate quark lines, the curly lines refer to the gauge bosons and the dashed lines denote the coupling of the axial singlet current  $J_5^\mu$ .

The net result is that the true global symmetry of the massless quantum theory is not  $U_L(3) \times U_R(3)$  but  $U_V(1) \times SU_L(3) \times SU_R(3)$ . The singlet vector symmetry,  $U_V(1)$  corresponds to baryon number conservation.

A similar situation emerges when external c-number flavor gauge fields are added to the massless QCD Lagrangian in order to further probe its structure. Then the so-called non-Abelian anomaly yields non-zero *covariant* derivatives of the  $SU_L(3) \times SU_R(3)$  currents proportional to certain combinations of the corresponding external gauge fields [28]. The non-Abelian anomaly will be noted to have important consequences for the theory of Skyrmions.

Although the true symmetry of massless three flavor QCD is  $SU_L(3) \times SU_R(3) \times U_V(1)$ , the resulting symmetry of the physical states of the theory is further reduced to  $SU_V(3) \times U_V(1)$  by the “spontaneous breakdown” mechanism. In this mechanism the vacuum state is not invariant under the full symmetry group. The massless QCD vacuum is characterized by a non-vanishing “condensate”  $\langle \bar{q}_u q_u + \bar{q}_d q_d + \bar{q}_s q_s \rangle \neq 0$ . Under an infinitesimal chiral transformation  $L = 1 + i \sum_{a=0}^{N_f^2-1} \epsilon_L^a \lambda^a / 2$  and  $R = 1 + i \sum_{a=0}^{N_f^2-1} \epsilon_R^a \lambda^a / 2$  the variation of this quark–bilinear is found to be

$$\delta(\bar{q}q) = i(\epsilon_L^a - \epsilon_R^a) \left( \bar{q}_L \frac{\lambda^a}{2} q_R - \bar{q}_R \frac{\lambda^a}{2} q_L \right) = (\epsilon_L^a - \epsilon_R^a) \bar{q} \frac{\lambda^a}{2} i\gamma_5 q. \quad (19)$$

Clearly the condensate is invariant only for the subgroup  $L = R$  in eq (17) which is a vector type transformation. This explains the physical  $SU_V(3) \times U_V(1)$  invariance. Note that the right hand side of eq (19) represents pseudoscalar objects. These are “zero mode” fluctuations of the above vacuum configuration and the corresponding massless pseudoscalar particles are designated Nambu–Goldstone bosons. Their scalar chiral “partners” – which would be degenerate in mass were it not for the spontaneous symmetry breakdown – are *not* constrained to be massless. This splitting of low-lying pseudoscalars and scalars expected from massless QCD seems in qualitative agreement with the experimental situation.

The  $SU_V(3) \times U_V(1)$  invariance (so-called “eightfold way”) of massless QCD is, of course, further broken when the effects of non-zero quark mass terms are included. For later purposes it is convenient to rewrite the quark mass terms as:

$$\mathcal{L}_{\text{QCD}}^{\text{mass}} = -\frac{m_u + m_d}{2} \bar{q} \mathcal{M} q \quad \text{with} \quad \mathcal{M} = \frac{2+x}{3} \mathbf{1} + y \lambda_3 + \frac{1-x}{\sqrt{3}} \lambda_8. \quad (20)$$

Here characteristic quark mass ratios are defined by

$$x = \frac{2m_s}{m_u + m_d}, \quad y = \frac{m_u - m_d}{m_u + m_d}. \quad (21)$$

In the limit  $y = 0$ , the theory possesses  $SU_V(2)$  or isospin invariance.

### 3.2 Effective Lagrangian of pseudoscalars

The simplest way to mock up low energy QCD is to employ a  $3 \times 3$  matrix field  $M_{ij}$  which transforms under the chiral group in the same way as the bilinear quark combination  $\bar{q}_{jR}q_{iL}$ . It has the decomposition  $M = S + iP$  into hermitian scalar and pseudoscalar components. A chirally invariant Lagrangian is

$$\mathcal{L} = \frac{1}{2} \text{tr}(\partial_\mu M \partial^\mu M^\dagger) - V(M, M^\dagger), \quad (22)$$

where the potential  $V$  is a function of invariants like  $\text{tr}(MM^\dagger)$ ,  $\text{tr}(MM^\dagger MM^\dagger)$  etc. Spontaneous breakdown to  $SU_V(3)$  is implemented by choosing  $V$  to have a minimum such that  $\langle M \rangle = \text{const.} \times \mathbf{1}$ . Then the scalar fields  $S$  become massive and can be “integrated out” by imposing a chirally invariant constraint [10]. This represents a transition from the linear to the non-linear sigma model. Formally we may use the “polar decomposition” of the matrix  $M$  into unitary and hermitian factors  $M = HU$ . Setting  $H \rightarrow \text{const.} \times \mathbf{1}$  then results in eq (1) again.

In principle, a scenario of this sort can be derived by adding a term like  $\bar{q}_L M q_R + h.c.$  to the QCD Lagrangian and then integrating out the quark fields<sup>2</sup>. As a result one is left with a complicated action functional for  $M$  and, by further eliminating  $H$  as above, for  $U$ . Note that  $U$  inherits the chiral transformation property of the quark bilinear:

$$U \longrightarrow LUR^\dagger. \quad (23)$$

It is this identification of the transformation properties which provides the important link of the effective chiral theory to QCD since it in particular implies that the (Noether) currents must be identified. In turn, matrix elements of these currents will yield the static properties of hadrons.

Of course, in the chiral limit we demand the effective meson theory to be strictly invariant under the transformation (23). Since  $U^\dagger U = 1$  only derivative terms can appear and the leading one is just the non-linear  $\sigma$  model (1). Clearly also the Skyrme term (5) is invariant under (23).

At this point one essential ingredient is still missing to ensure that the chiral field,  $U = \exp(i\Phi)$  describes pseudoscalar fields,  $\Phi$ . This requirement demands the parity transformation

$$U \xrightarrow{\text{parity}} U^\dagger. \quad (24)$$

However, it is straightforward to verify that the Skyrme model Lagrangian is invariant under (24) and  $\mathbf{r} \rightarrow -\mathbf{r}$  separately. On the level of the equations of motion we can easily break this unwanted extra symmetry by adding a term which contains the Levi-Civita tensor. With  $\alpha_\mu = (\partial_\mu U)U^\dagger$  we write

$$\frac{f_\pi^2}{2} \partial_\mu \alpha^\mu + \dots + 5\lambda \epsilon_{\mu\nu\rho\sigma} \alpha^\mu \alpha^\nu \alpha^\rho \alpha^\sigma = 0, \quad (25)$$

where the ellipsis refers to the contributions from the Skyrme term (5) which have the same symmetries as those from the non-linear  $\sigma$  term (1). Unfortunately, the additional term cannot be easily incorporated in the effective Lagrangian since it does not correspond to the variation of a local term. Witten suggested [30] to include it at the level of the action because the variation of

$$\Gamma = \lambda \int_{M_5} \text{tr} \epsilon_{\mu\nu\rho\sigma\tau} \alpha^\mu \alpha^\nu \alpha^\rho \alpha^\sigma \alpha^\tau \quad (26)$$

<sup>2</sup>In the case of QCD this seems to be impractical. However, the simpler Nambu–Jona–Lasinio [9] model for the quark flavor dynamics nicely exemplifies how a meson functional can be constructed by integrating out quark degrees of freedom [29].

and the use of Stoke's theorem yields the desired term in the equation of motion (25) provided the boundary of the five dimensional manifold,  $M_5$  is taken to be Minkowski space, *i.e.*  $\partial M_5 = M_4$ . The choice of  $M_5$  is not unique because its complement has the same boundary. In order to nevertheless have a unique action the constant  $\lambda = \frac{-in}{240\pi^2}$ ,  $n \in \mathbb{Z}$  must be quantized<sup>3</sup>. It is interesting to study the physical relevance of (26). Expanding in the meson fields  $\Phi$  and employing again Stoke's theorem reveals that it describes processes with at least five different pseudoscalars<sup>4</sup> like  $K^+K^- \rightarrow \pi^+\pi^0\pi^-$ . As such processes were first discussed by Wess and Zumino [31] who essentially found a power series expression for (26), the term is commonly named after them<sup>5</sup>.

There are further important consequences of the Wess–Zumino term (26) which can be read off after generalizing it so its variation with respect to external (electro–weak) gauge transformations [30, 32] yields the non–Abelian anomaly [28]. After appropriately including the corresponding gauge boson fields two striking features are observed:

- (i) A contact interaction for the decay  $\pi^0 \rightarrow \gamma\gamma$  is contained in the gauged Wess–Zumino action. On the quark level this process is described by the ABJ anomaly involving the diagrams of figure 1 with the external lines representing photons. Identifying that result with the Wess–Zumino term requires setting  $n = N_C$ , *i.e.* the Wess–Zumino term is proportional to the number of colors.
- (ii) The linear coupling to the  $U_V(1)$  gauge boson represents the baryon number current. Indeed it turns out that this current is identical to the topological current  $B_\mu$  in eq (3).

The mocking up of the effects of the  $U_A(1)$  anomaly in QCD involves the  $SU(3)$  singlet pseudoscalar particle  $\eta'$  and will be discussed later when we treat the *proton spin puzzle*.

We must also take account of the effects of the finite quark mass terms (20). These transform according to the chiral  $SU_L(3) \times SU_R(3)$  representation:  $\mathbf{3} \times \mathbf{3}^* + \mathbf{3}^* \times \mathbf{3}$ . Note that the matrix  $\mathcal{M} = \mathcal{M}^\dagger$  (neglecting the possibility of strong CP violation) may be considered a “spurion” for this transformation property. Then the minimal symmetry breaking piece of the effective Lagrangian reads

$$\mathcal{L}_{SB} = \text{tr} \{ \mathcal{M} [ -\beta' (\partial_\mu U \partial^\mu U^\dagger U + U^\dagger \partial_\mu U \partial^\mu U^\dagger) + \delta' (U + U^\dagger - 2) ] \}, \quad (27)$$

where  $\beta'$  and  $\delta'$  are two numerical parameters. The  $\delta'$  term is required to split the pseudoscalar meson masses while the  $\beta'$  term is required to split the pseudoscalar “decay constants”. The decay constants  $f_a$  are defined from the axial vector matrix elements  $\langle 0 | j_{5,\mu}^a | \phi_a, p \rangle = i f_a p_\mu$ .

Working in the isospin invariant limit, the parameters  $\beta'$ ,  $\delta'$  and  $x$  can then be extracted from the knowledge of meson properties [33],

$$m_\pi^2 = \frac{4}{f_\pi^2} \delta', \quad m_K^2 = \frac{4}{f_K^2} \delta' (1+x) \quad \text{and} \quad \left( \frac{f_K}{f_\pi} \right)^2 = 1 + \frac{4}{f_\pi^2} \beta' (1-x). \quad (28)$$

This represents the essential input when discussing the Skyrme model for three flavors. To sum up, the Lagrangian of only pseudoscalars which we shall use for discussing Skyrmions consists of the sum of (1), (5), (26) and (27).

In the chiral perturbation theory approach [12] essentially the most general chirally invariant Lagrangian is written down and ordered in powers of  $\partial\partial \sim \mathcal{M}$ . For example, the leading terms are eq (1) and the  $\delta'$  term of eq (27). The next-to-leading terms include:

$$\begin{aligned} & [\text{tr}(\partial_\mu U \partial^\mu U^\dagger)]^2, \text{tr}(\partial_\mu U \partial_\nu U^\dagger) \text{tr}(\partial^\mu U \partial^\nu U^\dagger), \\ & \text{tr}(\partial_\mu U \partial^\mu U^\dagger \partial_\nu U \partial^\nu U^\dagger), \text{tr}(\partial_\mu U \partial^\mu U^\dagger) \text{tr}(\mathcal{M}(U + U^\dagger)), \text{tr}(\partial_\mu U \partial^\mu U^\dagger \mathcal{M}(U + U^\dagger)), \\ & [\text{tr}(\mathcal{M}(U + U^\dagger))]^2, [\text{tr}(\mathcal{M}(U - U^\dagger))]^2, \text{tr}(\mathcal{M}U^\dagger \mathcal{M}U^\dagger + \mathcal{M}U \mathcal{M}U), \end{aligned} \quad (29)$$

<sup>3</sup>The reader is referred to the literature [30] to see the analogy to Dirac's quantization of the magnetic monopole.

<sup>4</sup>For that reason the term (26) vanishes in the case of two flavors which only has four different pseudoscalar fields.

<sup>5</sup>Some authors refer to the Wess–Zumino term in its gauged form *i.e.* with external sources. Unless otherwise noted we will always understand the Wess–Zumino term to be (26).

each with its own coupling constant. Note that the combinations of terms which can not be manipulated (by use of various matrix identities) to become a single trace are suppressed in the  $1/N_C$  expansion. This procedure also entails absorbing the divergent parts of loop corrections in the coefficients of the listed terms. The result is a joint power series in energy and the quark masses which can be expected to be very accurate quite near the  $\pi\pi$  threshold in the case of pion-pion scattering for example. But going higher in energy is very difficult in this scheme. Furthermore it seems that many of the coefficients mainly simulate the low energy effects of vector meson exchanges. We shall also work with a meson Lagrangian which includes the vector particles directly. This may be thought of as a start on the approach of constructing the leading (Born) term of the  $1/N_C$  expansion, which should include mesons of all spins.

### 3.3 Effective Lagrangian of pseudoscalars and vectors

We will follow the so-called massive Yang-Mills approach [32] for introducing the vector meson nonet into the Lagrangian of pseudoscalars in a chirally invariant manner. In this approach both vector and axial vector fields are formally introduced as gauge fields (yielding invariance under *local* chiral transformations). Then globally invariant mass-type terms which break the local chiral invariance are included. Finally, as in the transition from the linear to the non-linear sigma model discussed in section 3.1 above, the (heavier) axial vector mesons are eliminated by a chirally invariant constraint.

We introduce two multiplets with spin one,  $A_\mu^L$  and  $A_\mu^R$  which we demand to transform under (17) as left- and right-handed fields, respectively,

$$A_\mu^L \longrightarrow L \left( A_\mu^L + \frac{i}{g} \partial_\mu \right) L^\dagger \quad \text{and} \quad A_\mu^R \longrightarrow R \left( A_\mu^R + \frac{i}{g} \partial_\mu \right) R^\dagger. \quad (30)$$

This allows us to define a covariant derivative for the chiral field and field tensors,

$$D_\mu U = \partial_\mu U - ig A_\mu^L U + ig U A_\mu^R, \quad (31)$$

$$F_{\mu\nu}^{L,R} = \partial_\mu A_\nu^{L,R} - \partial_\nu A_\mu^{L,R} - ig [A_\mu^{L,R}, A_\nu^{L,R}], \quad (32)$$

which transform homogeneously under (17). The chirally invariant terms with a minimal number of derivatives read

$$\text{tr} [(D_\mu U)^\dagger D^\mu U], \quad \text{tr} [F_{\mu\nu}^{L,R} F^{L,R,\mu\nu}] \quad \text{and} \quad \text{tr} [F_{\mu\nu}^L U F^{R,\mu\nu} U^\dagger]. \quad (33)$$

In addition we can have mass-type terms for the vector mesons

$$\text{tr} [A_\mu^L A^{L,\mu} + A_\mu^R A^{R,\mu}] \quad \text{and} \quad \text{tr} [A_\mu^L U A^{R,\mu} U^\dagger], \quad (34)$$

which are still invariant under global chiral transformations. Of course, many more terms with higher derivatives could be written down at the expense of more undetermined parameters. Now, it is our aim to construct an effective model for the vector mesons only; at present we are not interested in the axial-vector mesons. We have to find a mechanism to eliminate the latter without violating the chiral symmetry. This can be accomplished by choosing a special “gauge” for the vector fields  $A_\mu^{L,R}$ ,

$$\tilde{A}_\mu^L = \xi \left( \rho_\mu + \frac{i}{g} \partial_\mu \right) \xi^\dagger \quad \text{and} \quad \tilde{A}_\mu^R = \xi^\dagger \left( \rho_\mu + \frac{i}{g} \partial_\mu \right) \xi. \quad (35)$$

Here  $\rho_\mu$  is a matrix field with  $N_f^2$  components. For example, in the case of two flavors it includes both the  $\rho$  and  $\omega$  mesons via  $\rho_\mu = \rho_\mu \cdot \tau + \omega_\mu \mathbf{1}$ . In the case of three flavors, this matrix field is supplemented by the  $K^*$  and  $\phi$  mesons. Most importantly we have introduced the “square root”,  $\xi$  of the chiral field,  $U = \xi\xi$  which yields the chirally invariant relation

$$\tilde{A}_\mu^L = U \left( \tilde{A}_\mu^R + \frac{i}{g} \partial_\mu \right) U^\dagger. \quad (36)$$

It is actually this so-called unitary constraint which eliminates the axial-vector fields in favor of the vector fields  $\rho$  without spoiling chiral symmetry.

It is interesting to study the behavior of the  $\rho$  meson under chiral transformations. To start off, we recognize that the transformation of  $\xi$  introduces the matrix  $K$  which is defined by [34]

$$\xi \longrightarrow L\xi K^\dagger \stackrel{!}{=} K\xi R^\dagger. \quad (37)$$

Clearly this leaves the transformation law of the chiral field (23) unchanged. Note that in general the matrix  $K$  is a position dependent quantity because of  $\xi$ . Demanding now the symmetry transformation

$$\rho_\mu \longrightarrow K \left( \rho_\mu + \frac{i}{g} \partial_\mu \right) K^\dagger \quad (38)$$

causes the fields  $\bar{A}^{L,R}$  to transform exactly like left- and right-handed vector fields.

Within the unitary gauge the various terms listed above in (33) and (34) are no longer independent. Introducing the homogeneously transforming combinations

$$p_\mu = \partial_\mu \xi \xi^\dagger + \xi^\dagger \partial_\mu \xi \quad \text{and} \quad R_\mu = \rho_\mu + \frac{i}{2g} (\partial_\mu \xi \xi^\dagger - \xi^\dagger \partial_\mu \xi) \quad (39)$$

the terms up to two derivatives can be combined to a chirally invariant Lagrangian of vectors (and pseudoscalars)

$$\mathcal{L}_{\text{VM}} = \text{tr} \left[ -\frac{1}{4} f_\pi^2 p_\mu p^\mu - \frac{1}{2} F_{\mu\nu}(\rho) F^{\mu\nu}(\rho) + m_\rho^2 R_\mu R^\mu \right], \quad (40)$$

where we have used the fact that the coefficient of the term quadratic in the  $\rho$  meson field is related to the vector meson mass  $m_\rho = 770\text{MeV}$ . Upon expanding the square-root field  $\xi$  in powers of the pseudoscalar field, one finds that the Lagrangian (40) contains the  $\rho\pi\pi$  coupling,

$$\mathcal{L}_{\rho\pi\pi} = \frac{m_\rho^2}{2g f_\pi^2} \rho_\mu \cdot (\pi \times \partial^\mu \pi), \quad (41)$$

which can be utilized to fix the coupling constant  $g \approx 5.6$  from the known decay-width of the process  $\rho \rightarrow \pi\pi$ .

Terms which involve the Levi-Civita tensor  $\epsilon_{\mu\nu\rho\sigma}$  are also of great interest for the Skyrme model. For their presentation it is most useful to introduce the notation of differential forms:  $A^R = A_\mu^R dx^\mu$ ,  $d = \partial_\mu dx^\mu$ , etc. . Since the left- and right-handed "gauge fields" are related via the unitary constraint (36) the number of linearly independent terms, which transform properly under chiral transformation as well as parity and charge conjugation, is quite limited [32]

$$A^L \alpha^3, \quad dA^L \alpha A^L - A^L \alpha dA^L + A^L \alpha A^L \alpha, \quad 2(A^L)^3 \alpha + \frac{i}{g} A^L \alpha A^L \alpha. \quad (42)$$

For convenience we have again made use of  $\alpha_\mu = (\partial_\mu U)U^\dagger = \xi p \xi^\dagger$ . Of course, including these terms in the model Lagrangian will introduce three more parameters:  $\gamma_1, \gamma_2$  and  $\gamma_3$ . A suitable presentation of this part of the action is given in terms of  $p$  and  $R$  (employing again the notation of differential forms)

$$\Gamma_\epsilon = \Gamma_{\text{WZ}} + \int_{M_4} \text{tr} \left( \frac{1}{6} \left[ \gamma_1 + \frac{3}{2} \gamma_2 \right] R p^3 - \frac{i}{4} g \gamma_2 F(\rho) [pR - Rp] - g^2 [\gamma_2 + 2\gamma_3] R^3 p \right), \quad (43)$$

where  $\Gamma_{\text{WZ}}$  is given in (26). In ref [35] two of the three unknown constants,  $\gamma_{1,2,3}$  were determined from purely strong interaction processes like  $\omega \rightarrow 3\pi$ . Defining  $\tilde{h} = -2\sqrt{2}\gamma_1/3$ ,  $\tilde{g}_{VV\phi} = g\gamma_2$  and  $\kappa = \gamma_3/\gamma_2$  the central values  $\tilde{h} = \pm 0.4$  and  $\tilde{g}_{VV\phi} = \pm 1.9$  were found. Within experimental

uncertainties (stemming from the errors in the  $\omega - \phi$  mixing angle) these may vary in the range  $\tilde{h} = -0.15, \dots, 0.7$  and  $\tilde{g}_{VV\phi} = 1.3, \dots, 2.2$  subject to the condition  $|\tilde{g}_{VV\phi} - \tilde{h}| \approx 1.5$ . The third parameter,  $\kappa$  could not be fixed in the meson sector. From studies [36] of nucleon properties in the two flavor model it was argued that  $\kappa \approx 1$  represents a reasonable choice.

The sum of the space integral of (40) and (43) comprise the chirally invariant part of the effective action we shall use for discussing the soliton in the vector meson model. Note that the second piece of (43) can be gauged with external fields [35] so as to make no contribution to the non-Abelian anomaly. The first piece  $\Gamma_{WZ}$  then correctly supplies the non-Abelian anomaly. Furthermore the second piece of (43) stabilizes the soliton without the need for including the Skyrme term (5).

We must still include the effects of symmetry breaking due to finite quark masses in the vector meson system. To leading order in the symmetry breaking, an appropriate term which behaves properly under chiral transformations, can be constructed by analogy to the last expression in (34)

$$-\alpha' \text{tr} [\mathcal{M} (A_\mu^L U A^{R\mu} + A_\mu^R U^\dagger A^{L\mu})]. \quad (44)$$

This leading contribution not only distinguishes between the  $\rho$  and  $K^*$  masses but also contributes to the different decay constants of the pseudoscalar mesons via the unitary gauge (35). The reader may consult ref [33] for recent discussion of higher order symmetry breaking terms.

We would like to end this section on including vector mesons by noting that the same Lagrangian is obtained within the so-called hidden gauge approach [37], once the same symmetries are required. This shows that these two approaches are in fact identical.

### 3.4 Other aspects

We expect that baryons should appear as solitons of the large  $N_C$  effective meson Lagrangian for any number of flavors  $N_f$ . In the case of three (or more) light flavors the Wess–Zumino term guarantees, as discussed in section 3.1, that the baryon number (3) is obtained in a self-contained manner from the Lagrangian. This can be used to check that the soliton indeed has the correct baryon number.

Now in the two flavor case, the same kind of soliton solution exists. However the Wess–Zumino term vanishes identically so we cannot similarly check its baryon number in a self-contained way. The situation is even more peculiar for  $N_f = 1$ . There the Skyrme model represents a mapping  $S^3 \rightarrow S^1$  which does not contain topologically stable configurations. However, we are not forced to use an effective Lagrangian of the same form. In this case it is probably more realistic to construct the Lagrangian by including isoscalars like the spin-0  $\sigma$ -field and the spin-1  $\omega$ -field. Such a Lagrangian might have a soliton solution (not necessarily topological) but a check of its baryon number may also not be available in a self-contained way. These examples seem to indicate that the form of the relevant effective Lagrangian may have a non-trivial  $N_f$  dependence (at least for small  $N_f$ ).

Another interesting question, related in the sense of understanding whether physical features of the solitons can be traced to particular pieces of the effective Lagrangian, concerns the *stabilization* of the soliton. In section 2 we noted that the Skyrme term (5) was introduced precisely for this purpose. There is an often mentioned “derivation” of this term from the piece of the vector meson Lagrangian (40) above which goes as follows. In a large mass expansion,  $m_\rho \rightarrow \infty$  the equation of motion for the vector meson field simply becomes  $R_\mu = 0$ . Substituting this into the remainder of the vector meson Lagrangian (40)

$$F_{\mu\nu}(\rho) \longrightarrow F_{\mu\nu} \left( \frac{-i}{2g} [\partial_\mu \xi \xi^\dagger - \xi^\dagger \partial_\mu \xi] \right)$$

yields exactly the Skyrme term (5) with the identification  $g = e$ . Although the numbers 5.6 and 4.75 are in reasonable agreement there is one caveat to this appealing derivation of the Skyrme term. While the Skyrme model does yield stable solitons, however, for arbitrary large but finite  $m_\rho$  the model (40) does not contain stable soliton solutions. Thus one seems to have achieved stabilization merely by approximating a model in which stabilization does not exist. Clearly



we have not obtained a “physical origin” for the stabilization mechanism. As mentioned in the previous section, the second piece of (43) stabilizes the soliton in the vector meson Lagrangian without a need for the Skyrme term. It is also possible that, as in the case of the s-wave ground state of hydrogen, stability is achieved at the quantum, rather than at the classical, level. Several investigations of this possibility have been made [38] based on just the non-linear sigma model term (1), although an assumption on the allowed chiral profiles seems to be required.

## 4 The Skyrme model with three flavors

It is well established [7] that the neutron ( $n$ ) and proton ( $p$ ) belong to a multiplet with six other members (the iso-singlet  $\Lambda$ , the iso-doublet  $\Xi$  and the iso-triplet  $\Sigma$ ). To try to understand  $n$  and  $p$  alone is to look at only a small piece of a large picture. Thus we must consider the three flavor generalization of the treatment in section 2. First (in the present section) we shall consider the Lagrangian of pseudoscalars alone, discussed in section 3.2. The new features arise from the inclusion of flavor  $SU(3)$  symmetry breaking terms (see (27) together with (20) and (21)) as well as the Wess–Zumino term (26). Both of these features involve non-trivial extensions of the formalism and interesting “physics”.

The first step towards including the strangeness degrees of freedom is to actually take the chiral field to be a  $U(1) \otimes SU(3)$  matrix. To be precise, the three flavor chiral field is defined as

$$U(x) = \exp \left( i \frac{\sqrt{2}}{\sqrt{3}f_\pi} \eta_0 \right) \exp(i\Phi). \quad (45)$$

While the singlet field  $\eta_0$  is separated the matrix field  $\Phi$  now not only contains the pion degrees of freedom but also the kaons and the non-singlet component of the  $\eta$  fields,

$$\Phi = \sum_{a=1}^8 \frac{\sqrt{2}}{f_\pi} \phi^a \lambda^a = \begin{pmatrix} \frac{1}{\sqrt{2}}\pi^0 + \frac{1}{\sqrt{6}}\eta_8 & \pi^+ & K^+ \\ \pi^- & -\frac{1}{\sqrt{2}}\pi^0 + \frac{1}{\sqrt{6}}\eta_8 & K^0 \\ K^- & \bar{K}^0 & -\frac{2}{\sqrt{6}}\eta_8 \end{pmatrix}. \quad (46)$$

Here  $\lambda^a$  denote the Gell–Mann matrices. Note that in the presence of derivative-type symmetry breakers (*e.g.* the  $\beta'$  term in (27)) the normalization of the fields gets shifted; the “physical” fields are gotten by multiplying the fields above by some constants as  $Z_\pi\pi^+$ ,  $Z_K K^+$ , etc.; similarly the physical decay constants are  $Z_\pi f_\pi = 93\text{MeV}$ ,  $Z_K f_K \approx 113\text{MeV}$ , etc. For the  $Z$ ’s we have

$$Z_\pi = \left( 1 - \frac{8}{f_\pi^2} \beta' \right)^{\frac{1}{2}}, \quad Z_K = \left( 1 - \frac{4}{f_\pi^2} (1+x) \beta' \right)^{\frac{1}{2}} \quad \text{etc.} \quad (47)$$

We clearly need a suitable generalization of the Skyrme *ansatz* (7). It turns out that it is correct to just embed the  $SU(2)$  hedgehog in the  $SU(3)$  matrix. Flavor symmetry breaking implies that field configurations which have non-zero strangeness possess a classical energy which (at least in the unit baryon number sector) is larger than that of a zero strangeness configuration. Thus we choose the embedding:

$$U_0(\mathbf{r}) = \left( \begin{array}{c|c} \exp(i\boldsymbol{\tau} \cdot \hat{\mathbf{r}} F(r)) & \begin{smallmatrix} 0 \\ 0 \end{smallmatrix} \\ \hline 0 & 1 \end{array} \right). \quad (48)$$

Hence the classical energy will not be modified and the soliton profile,  $F(r)$  is that in figure 2.1. The effects of the strange degrees of freedom are hence visible when states with baryon quantum numbers are generated via the collective coordinate approach.

The collective coordinate matrix  $A(t)$  defined in eq (9) is now taken from  $SU(3)$  and in analogy to eq (11), now leads to eight angular velocities,

$$A^\dagger(t) \frac{d}{dt} A(t) = \frac{i}{2} \sum_{a=1}^8 \lambda^a \Omega_a. \quad (49)$$

In addition to the angular velocities  $\Omega_a$  the adjoint representation

$$D_{ab} = \frac{1}{2} \text{tr} (\lambda_a A \lambda_b A^\dagger) \quad (50)$$

of the collective rotations,  $A(t)$  will be important, in particular in the context of flavor symmetry breaking.

Substituting  $U = A(t)U_0(r)A^\dagger(t)$  into the pseudoscalar Lagrangian of section 3.2 without the symmetry breaker gives rise after a spatial integration to the collective Lagrangian

$$L_{\text{Skyrme}}(A, \dot{A}) + L_{\text{WZ}}(A, \dot{A}) = \frac{1}{2} \alpha^2 \sum_{i=1}^3 \Omega_i^2 + \frac{1}{2} \beta^2 \sum_{\alpha=4}^7 \Omega_\alpha^2 - \frac{N_C B}{2\sqrt{3}} \Omega_8 - E_{\text{cl}}. \quad (51)$$

The  $SU(2)$  moment of inertia  $\alpha^2$  remains unchanged while the moment of inertia  $\beta^2$  for rotations into the strange directions is a new functional of the pseudoscalar fields. The fact that the eighth component of the angular velocity vector does not appear quadratically in (51) is a consequence of  $[U_0, \lambda_8] = 0$ . The term proportional to  $B\Omega_8$ , where  $B$  is the baryon number arises from  $\Gamma_{\text{WZ}}$ . In order to obtain it we make use of the separation [39]

$$\Gamma_{\text{WZ}}[U] = \Gamma_{\text{WZ}}[U_0] - \frac{iN_C}{48\pi^2} \int d^4x \text{tr} \left\{ \left[ (U_0^\dagger dU_0)^3 + (U_0 dU_0^\dagger)^3 \right] (A^\dagger dA) \right\}, \quad (52)$$

where, again, Stoke's theorem has been employed. As  $U_0$  is static we have  $\Gamma_{\text{WZ}}[U_0] = 0$  and the remainder becomes a local object which is straightforwardly evaluated.

#### 4.1 Quantization of the three flavor collective Lagrangian

In order to quantize the three flavor Lagrangian (51) we require the operators for spin and flavor as Noether charges. As a consequence of the hedgehog structure, the infinitesimal change under spatial rotations can be written as a derivative with respect to  $\Omega$

$$[\mathbf{r} \times \partial, U(\mathbf{r}, t)] = \frac{\partial \dot{U}(\mathbf{r}, t)}{\partial \Omega}. \quad (53)$$

By the Noether construction this leads to the spin operator  $\mathbf{J} = \partial L(A, \Omega_a) / \partial \Omega$ . The quantization of the “ $SU(3)$  rigid top” proceeds by generalizing this result to the so-called right generators

$$R_a = -\frac{\partial}{\partial \Omega_a} (L_{\text{Skyrme}} + L_{\text{WZ}}) = \begin{cases} -\alpha^2 \Omega_a = -J_a, & a=1,2,3 \\ -\beta^2 \Omega_a, & a=4,\dots,7 \\ \frac{N_C B}{2\sqrt{3}}, & a=8 \end{cases}. \quad (54)$$

The quantization prescription then demands the commutation relation  $[R_a, R_b] = -if_{abc}R_c$  with  $f_{abc}$  being the antisymmetric structure constants of  $SU(3)$ . Explicit expressions for these generators in terms of an “Euler-angle” parameterization of  $A$  are presented in ref [40]. The so-called left generators, which are defined by the rotation  $L_a = D_{ab}R_b$ , satisfy the commutation relations  $[L_a, L_b] = if_{abc}L_c$ . They provide the isospin,  $I_i = L_i$  ( $i = 1, 2, 3$ ) and hypercharge,  $Y = 2L_8/\sqrt{3}$  operators.

The generator  $R_8$  is linearly connected to the so-called right hypercharge  $Y_R = 2R_8/\sqrt{3} = 1$  for  $B = 1$  and  $N_C = 3$ . In analogy to the Gell-Mann Nishijima relation a right charge

$$Q_R = -J_3 + \frac{Y_R}{2} \quad (55)$$

may be defined. Completing the analogy we note that the eigenvalues of  $Q_R$  are  $0, \pm 1/3, \pm 2/3, \pm 1, \dots$ . Hence for  $Y_R = 1$  the relation (55) can only be fulfilled when the eigenvalue of  $J_3$  is half-integer. This yields the important conclusion that the  $SU(3)$  model describes fermions. *A priori* this is not expected since the starting point has been an effective model of bosons. This discussion

can be generalized to arbitrary  $N_C$  showing that the Skyrmin describes fermions when  $N_C$  is odd and bosons when  $N_C$  is even [30]. This, of course, is expected from considering baryons as being composed of  $N_C$  quarks. We conclude that the proper incorporation of the anomaly structure of QCD leads to the desired spin-statistics relation.

## 4.2 Flavor symmetry breaking and baryon spectrum

For a realistic treatment of baryon states in the space of the collective coordinates we have to supplement the collective Lagrangian by the flavor symmetry breaking pieces associated with (27). Substituting the flavor rotating hedgehog yields the symmetry breaking piece in the collective Lagrangian,

$$L_{\text{SB}} = -\frac{1}{2}\gamma(1 - D_{88}) \quad (56)$$

with the coefficient,  $\gamma$  being linear in the symmetry breaking parameter  $1 - x$ , *i.e.*

$$\gamma = \frac{32\pi}{3}(x - 1) \int dr \{ \delta' r^2 (1 - \cos F) - \beta' \cos F (F'^2 r^2 + 2 \sin^2 F) \}. \quad (57)$$

$D_{88}(A)$  is defined in eq (50). Putting pieces (51) and (56) together, the Hamiltonian for the collective coordinates is obtained as the Legendre transform  $H = -\sum_{a=1}^8 R_a \Omega_a - L$

$$H(A, R_a) = E_{\text{cl}} + \frac{1}{2} \left[ \frac{1}{\alpha^2} - \frac{1}{\beta^2} \right] \mathbf{J}^2 + \frac{1}{2\beta^2} C_2(SU(3)) - \frac{3}{8\beta^2} + \frac{1}{2}\gamma(1 - D_{88}) \quad (58)$$

for  $B = 1$  and  $N_C = 3$ . The constraint  $R_8 = \frac{\sqrt{3}}{2}$ , which yielded the spin-statistics relation, commutes with  $H$  permitting one to substitute this value. The term involving  $\sum_{\alpha=4}^7 R_\alpha^2$  has been re-expressed by introducing the quadratic Casimir operator of  $SU(3)$ ,  $C_2(SU(3)) = \sum_{a=1}^8 R_a^2$ . The standard  $SU(3)$  representations are eigenstates of  $C_2(SU(3))$  with eigenvalues  $\mu$ . For example, the octet representation **8** has  $\mu_8 = 3$  while  $\mu_{10} = \mu_{\overline{10}} = 6$  and  $\mu_{27} = 8$ . These representations diagonalize<sup>6</sup> the collective Hamiltonian in the absence of symmetry breaking,  $\gamma = 0$ .

Now consider the full collective Hamiltonian including the symmetry breaking. It seems reasonable to assume these  $\gamma = 0$  eigenstates as a basis to diagonalize the full Hamiltonian. In a perturbative treatment up to third order in  $\gamma$  for the  $\frac{1}{2}^+$  baryons only the representations **8**,  **$\overline{10}$**  and **27** contribute [42]. For that reason the perturbative treatment is still simple, although one must go beyond leading order. In particular this implies that the nucleon is no longer a pure octet state but rather contains sizable admixture of the nucleon type states in higher dimensional representations,

$$|N\rangle = |N, \mathbf{8}\rangle + 0.0745\gamma\beta^2 |N, \overline{\mathbf{10}}\rangle + 0.0490\gamma\beta^2 |N, \mathbf{27}\rangle + \dots, \quad (59)$$

where the coefficients of the effective symmetry breaker  $\gamma\beta^2$  are computed from  $SU(3)$  Clebsch-Gordon coefficients [43]. The nucleon is seen to have a roughly 25% amplitude to contain the  **$\overline{10}$**  state.

Although this perturbative treatment provides a physical picture of the symmetry breaking effects it actually turns out that the full Hamiltonian (58) can be exactly diagonalized numerically. The important ingredient is that within a suitable ‘‘Euler-angle’’ representation of the rotations  $A$ , the symmetry breaker  $1 - D_{88}$  depends only on one of these eight angles. In each isospin channel the eigenvalue equation

$$[C_2(SU(3)) + \beta^2\gamma(1 - D_{88})] \Psi = \epsilon_{\text{SB}} \Psi \quad (60)$$

then reduces to a set of coupled ordinary differential equations which can be integrated numerically. Here we do not wish to discuss this approach in full detail; rather we refer the reader to the original

<sup>6</sup>The hedgehog structure of the classical configuration  $U_0$  constrains the permissible  $SU(3)$  irreducible representations to those which have at least one state with  $I = J$  [41].

work by Yabu and Ando [44] and exhaustive applications of this method involving the present authors [45, 46, 40]. Having obtained the eigenvalue  $\epsilon_{\text{SB}}$  the baryon masses are straightforwardly computed from

$$M_B = E + \frac{1}{2} \left( \frac{1}{\alpha^2} - \frac{1}{\beta^2} \right) J(J+1) - \frac{3}{8\beta^2} + \frac{1}{2\beta^2} \epsilon_{\text{SB}}. \quad (61)$$

As already mentioned this diagonalization procedure is equivalent to the perturbation expansion. For small enough symmetry breaking  $\beta^2\gamma$  even first order is sufficient. In that (unjustified) case the famous Gell-Mann-Okubo mass formulae [47, 7] holds exactly:

$$2(M_N + M_{\Xi}) = M_{\Sigma} + 3M_{\Lambda} \quad (62)$$

$$M_{\Omega} - M_{\Xi^*} = M_{\Xi^*} - M_{\Sigma^*} = M_{\Sigma^*} - M_{\Delta}. \quad (63)$$

Additional corrections [45] arise when we allow for non-zero classical  $K$ -meson fields to get induced by “rotations”  $\Omega_{\alpha}$  into the strange directions. These are energetically favorable since they maximize the strange moment of inertia  $\beta^2$ . With a parameterization

$$\begin{pmatrix} K^+ \\ K^0 \end{pmatrix} = W(r) \hat{r} \cdot \tau \begin{pmatrix} \Omega_4 - i\Omega_5 \\ \Omega_6 - i\Omega_7 \end{pmatrix}, \quad (64)$$

the radial function  $W(r)$  is determined from applying a variational principle to  $\beta^2$ . In principle one must enforce that the *ansatz* (64) has no overlap with any global rotation of the classical solution (48).

We adjust the only free parameter,  $e \approx 4$  to the mass differences of the low-lying  $\frac{1}{2}^+$  and  $\frac{3}{2}^+$  baryons. The resulting baryon spectrum is shown in table 1. Apparently the three flavor Skyrme

Table 1: The mass differences, which are obtained by exact diagonalization of the collective Hamiltonian (58), of the  $\frac{1}{2}^+$  and  $\frac{3}{2}^+$  baryons in the pseudoscalar model for  $e=4.0$  are compared to the experimental data. The values in parentheses are obtained by enforcing the zero overlap condition mentioned after (64) [40]. In that case the Skyrme parameter has slightly been readjusted to  $e=3.9$ . All data are in MeV.

Baryons	Model	Expt.
$\Lambda - N$	154 (163)	177
$\Sigma - N$	242 (264)	254
$\Xi - N$	366 (388)	379
$\Delta - N$	278 (268)	293
$\Sigma^* - N$	410 (406)	446
$\Xi^* - N$	544 (545)	591
$\Omega - N$	677 (680)	733

model reasonably accounts for the empirical mass differences. The original studies [48, 49, 50, 44] yielded far too low mass splittings between baryons of different strangeness for physically motivated parameters of the effective Lagrangian<sup>7</sup>. A major reason for the improvement is the fact that  $\gamma$  is significantly enlarged by including the effects associated with  $f_K \neq f_{\pi}$  [45]. It is also apparent from table 1 that enforcing the zero overlap condition for the induced kaon components can be compensated by a small variation of the Skyrme parameter,  $e$ . This indicates that possible double counting effects play only a minor role. It is interesting to remark that the mass differences for the  $\frac{1}{2}^+$  baryons deviate strongly from the predictions in leading order of the flavor symmetry breaking. This can easily be observed from the ratios

$$(M_{\Lambda} - M_N) : (M_{\Sigma} - M_{\Lambda}) : (M_{\Xi} - M_{\Sigma}) = 1 : 0.52 : 0.85 \quad (65)$$

<sup>7</sup>Many of these authors considered  $f_{\pi}$  as a free parameter fitted to the absolute values of the baryon masses. Without the  $\beta'$  term this yielded  $f_{\pi}$  as low as 25MeV [49].

which are in much better agreement with the experimental data (1:0.43:0.69) than the leading order result (1:1:0.5). Obviously the higher order contributions are important. This also indicates that the baryon wave-functions contain sizable admixture of higher dimensional  $SU(3)$  representations, *cf.* eq (59). Nevertheless the deviation from the Gell-Mann-Okubo relations (62) is only moderate, in particular the equal spacing among the  $\frac{3}{2}^+$  baryons is well reproduced. Finally we note that, as discussed in point (iii) of section 2, the absolute mass of the nucleon is also too high in the three flavor case. Again we must rely on the  $N_C^0$  corrections mentioned.

### 4.3 Electromagnetic properties of $\frac{1}{2}^+$ baryons

The value for the Skyrme parameter  $e = 4.0$  obtained from this best fit to the baryon mass differences is next employed to evaluate static properties of baryons within this model. In order to do so one first constructs the Noether currents associated with the symmetry transformation (23). A convenient method is to extend these global symmetries to local ones by introducing external gauge fields (*e.g.* the gauge fields of the electroweak interactions) into the total action *i.e.* (1), (5), (26) and (27). The Noether currents are then read off as the expressions which couple linearly to these gauge fields. This procedure is especially appropriate for the Wess-Zumino term (26) because this non-local term can only be made gauge invariant by a lengthy iterative procedure [30, 32]. The final form of the nonet ( $a = 0, \dots, 8$ ) vector ( $V_\mu^a$ ) and axial-vector ( $A_\mu^a$ ) currents reads [46] (for  $N_C = 3$ )

$$\begin{aligned} V_\mu^a(A_\mu^a) = & -\frac{i}{2}f_\pi^2 \operatorname{tr} \{ (\xi Q^a \xi^\dagger \mp \xi^\dagger Q^a \xi) p_\mu \} - \frac{i}{8e^2} \operatorname{tr} \{ (\xi Q^a \xi^\dagger \mp \xi^\dagger Q^a \xi) [p_\nu, [p_\mu, p_\nu]] \} \\ & - \frac{1}{16\pi^2} \epsilon^{\mu\nu\rho\sigma} \operatorname{tr} \{ (\xi Q^a \xi^\dagger \pm \xi^\dagger Q^a \xi) p_\nu p_\rho p_\sigma \} \\ & - i\beta' \operatorname{tr} \{ Q^a (\{UM + MU^\dagger, \alpha_\mu\} \mp \{MU + U^\dagger M, \beta_\mu\}) \}, \end{aligned} \quad (66)$$

where  $Q^a = (\frac{1}{3}, \frac{\lambda^1}{2}, \dots, \frac{\lambda^8}{2})$  denote the Hermitian nonet generators. The combination

$$Q^{\text{e.m.}} = \operatorname{diag} \left( \frac{2}{3}, -\frac{1}{3}, -\frac{1}{3} \right) = Q^3 + \frac{1}{\sqrt{3}} Q^8 \quad (67)$$

is of special interest because it enters the computation of the electromagnetic properties. The associated form factors of the  $\frac{1}{2}^+$  baryons ( $B$ ) are defined by

$$\langle B(p') | V_\mu^{\text{e.m.}} | B(p) \rangle = \bar{u}(p') \left[ \gamma_\mu F_1^B(q^2) + \frac{\sigma_{\mu\nu} q^\nu}{2M_B} F_2^B(q^2) \right] u(p), \quad q_\mu = p_\mu - p'_\mu. \quad (68)$$

Frequently it is convenient to introduce “electric” and “magnetic” form factors

$$G_E^B(q^2) = F_1^B(q^2) - \frac{q^2}{4M_B^2} F_2^B(q^2), \quad G_M^B(q^2) = F_1^B(q^2) + F_2^B(q^2). \quad (69)$$

Substituting the rotating hedgehog configuration into the defining equation of the currents (66) yields for the spatial components of the vector current<sup>8</sup>

$$\begin{aligned} V_i^a = & V_1(r) \epsilon_{ijk} x_j D_{ak} + \frac{\sqrt{3}}{2} B(r) \epsilon_{ijk} \Omega_j x_k D_{a8} + V_2(r) \epsilon_{ijk} x_j d_{d\alpha\beta} D_{a\alpha} \Omega_\beta \\ & + V_3(r) \epsilon_{ijk} x_j D_{88} D_{ak} + V_4(r) \epsilon_{ijk} x_j d_{d\alpha\beta} D_{8\alpha} D_{a\beta} + \dots, \end{aligned} \quad (70)$$

where

$$B(r) = \frac{-1}{2\pi^2} F' \frac{\sin^2 F}{r^2} \quad (71)$$

<sup>8</sup>The conventions are  $i, j, k = 1, 2, 3$  and  $\alpha, \beta = 4, \dots, 7$ .

Table 2: The electromagnetic properties of the baryons compared to the experimental data. The predictions of the Skyrme model are taken from ref [46].

$B$	$\mu_B(\text{n.m.})$		$r_M^2(\text{fm}^2)$		$r_E^2(\text{fm}^2)$	
	$e = 4.0$	Expt.	$e = 4.0$	Expt.	$e = 4.0$	Expt.
$p$	2.03	2.79	0.43	0.74	0.59	0.74
$n$	-1.58	-1.91	0.46	0.77	-0.22	-0.12
$\Lambda$	-0.71	-0.61	0.36	—	-0.08	—
$\Sigma^+$	1.99	2.42	0.45	—	0.59	—
$\Sigma^0$	0.60	—	0.36	—	-0.02	—
$\Sigma^-$	-0.79	-1.16	0.58	—	-0.63	—
$\Xi^0$	-1.55	-1.25	0.38	—	-0.15	—
$\Xi^-$	-0.64	-0.69	0.43	—	-0.49	—
$\Sigma^0 \rightarrow \Lambda$	-1.39	-1.61	0.48	—	—	—

is the baryon number density (3). The explicit form of the radial functions  $V_1(r), \dots, V_4(r)$  is given in appendix B of ref [46]. According to the quantization prescription (54) the angular velocities  $\Omega_a$  are replaced by their expressions in terms of the right generators  $R_a$  of  $SU(3)$ . Taking the Fourier transform of the resulting matrix elements allows one to identify the magnetic form factor in the Breit frame [51, 52]

$$G_M^B(q^2) = -8\pi M_B \int_0^\infty r^2 dr \frac{r}{|q|} j_1(r|q|) \left\{ V_1(r) \langle D_{e3} \rangle_B - \frac{1}{2\alpha^2} B(r) \langle D_{e8} R_8 \rangle_B \right. \\ \left. - \frac{1}{\beta^2} V_2(r) \langle d_{3\alpha\beta} D_{e\alpha} R_\beta \rangle_B + V_3(r) \langle D_{88} D_{e3} \rangle_B + V_4(r) \langle d_{3\alpha\beta} D_{e\alpha} D_{8\beta} \rangle_B \right\}. \quad (72)$$

Here the flavor index  $e$  refers to the “electromagnetic” direction (67). The magnetic moment corresponds to the magnetic form factor at zero momentum transfer  $\mu_B = G_M^B(0)$ . Similarly the electric form factor is given by Fourier transforming the time component of the electromagnetic current

$$G_E^B = 4\pi \int_0^\infty r^2 dr j_0(r|q|) \left\{ \frac{\sqrt{3}}{2} B(r) \langle D_{e3} \rangle_B + \frac{1}{\alpha^2} V_7(r) \langle D_{ei} R_i \rangle_B + \frac{1}{\beta^2} V_8(r) \langle D_{e\alpha} R_\alpha \rangle_B \right\}. \quad (73)$$

The two new radial functions  $V_7(r)$  and  $V_8(r)$  are listed in appendix B of ref [46] as well. Integrating  $V_7$  and  $V_8$  yields the moments of inertia,  $\alpha^2$  and  $\beta^2$ , respectively. Hence the electric charges are properly normalized. It should be remarked that the baryon matrix elements in the space of the collective coordinates are computed using the exact eigenstates of (60) and adopting the Euler-angle representations for the  $SU(3)$  generators [40]. The results for the magnetic moments and the radii

$$r_M^2 = -\frac{6}{\mu_B} \frac{dG_M^B(q^2)}{dq^2} \bigg|_{q^2=0} \quad \text{and} \quad r_E^2 = -6 \frac{dG_E^B(q^2)}{dq^2} \bigg|_{q^2=0} \quad (74)$$

are shown in table 2. As in the two flavor model [6] the isovector part of the magnetic moments is underestimated while the isoscalar part is reasonably well reproduced. Despite the fact that the flavor symmetry breaking is large for the baryon wave-functions, the predicted magnetic moments do not strongly deviate from the  $SU(3)$  relations [53]

$$\mu_{\Sigma^+} = \mu_p, \quad \mu_{\Sigma^0} = \frac{1}{2}(\mu_{\Sigma^+} + \mu_{\Sigma^-}), \quad \mu_{\Sigma^-} = \mu_{\Xi^-}, \\ 2\mu_\Lambda = -(\mu_{\Sigma^+} + \mu_{\Sigma^-}) = -2\mu_{\Sigma^0} = \mu_n = \mu_{\Xi^0} = \frac{2}{\sqrt{3}}\mu_{\Sigma^0\Lambda}. \quad (75)$$

A more elaborate treatment of the flavor symmetry breaking is necessary in order to accommodate the experimentally observed details of breaking the  $U$ -spin symmetry which *e.g.* causes the approximate identity  $\mu_{\Sigma^+} \approx \mu_p$  [54]. The moderate differences between the various magnetic radii  $r_M^2$  is a further hint that symmetry breaking effects are mitigated. The comparison with the available empirical data for the radii shows that the predictions turn out too small in magnitude (except for the neutron electric radius). This is a strong indication that essential ingredients are still missing in the model. In section 5 it will be explained that the effects, which are associated with vector meson dominance (VMD), will account for this deficiency. Nevertheless the overall picture gained for the electromagnetic properties of the  $\frac{1}{2}^+$  and  $\frac{3}{2}^+$  can at least be characterized as satisfactory, especially in view of the fact that the only free parameter of the model has been fixed beforehand.

#### 4.4 Effects of symmetry breaking on baryon matrix elements and strangeness in the nucleon

Theorists typically wish for symmetry breaking effects to be negligible (the so-called “spherical cow” approximation) but Nature says otherwise. This is very apparent in the case of low energy strong interactions (QCD). The Gell-Mann–Okubo mass formulae, which amount to applications of a Wigner–Eckart theorem for first order  $\lambda_8$  type symmetry breaking, furnish sum rules rather than a complete description. As we discussed in section 4.2, the Skyrme model provides a non-trivial playground for treating symmetry breaking. The Yabu–Ando equation (60) gives an exact (within the model) wave-function for each baryon state at any strength of the symmetry breaking parameter  $\gamma\beta^2$  ( $\propto$  underlying quark masses). The physical results vary smoothly with  $\gamma\beta^2$  (although higher quantum corrections would be expected to give weak non-analytic corrections).

The physical interpretation of symmetry breaking in the model may be seen from eq (59). The higher  $SU(3)$  representation components in the baryon wave-function can only emerge in a quark framework by having quark–antiquark pairs present in addition to the three “valence” quarks. Clearly such effects would be difficult to treat in the non-relativistic quark model approach. On the other hand, it should be recognized that the Skyrme model approach is based on a collective semi-classical treatment.

In the last few years there has been a greatly renewed interest in the study of symmetry breaking effects for ordinary nucleons. This was stimulated by new experiments on polarized lepton deep inelastic scattering off nucleons [55] which seem to indicate that the pure valence quark picture of the nucleon has serious drawbacks. The Skyrme model has the advantage of giving a simple and roughly accurate quantitative explanation of these experiments. In detail one needs the strangeness conserving proton matrix elements

$$\langle P(\mathbf{p}') | \bar{q}_i \gamma_\mu \gamma_5 q_i | P(\mathbf{p}) \rangle = \bar{u}(\mathbf{p}') \left[ \gamma_\mu H_i(q^2) + \frac{q_\mu}{2M_p} \tilde{H}_i(q^2) \right] \gamma_5 u(\mathbf{p}) \quad (76)$$

for this discussion. Of related interest are the flavor changing matrix elements

$$\begin{aligned} \langle B'(\mathbf{p}') | V_\mu^\alpha | B(\mathbf{p}) \rangle &= \bar{u}(\mathbf{p}') [\gamma_\mu g_V(q^2) + \dots] u(\mathbf{p}) , \\ \langle B'(\mathbf{p}') | A_\mu^\alpha | B(\mathbf{p}) \rangle &= \bar{u}(\mathbf{p}') [\gamma_\mu \gamma_5 g_A(q^2) + \dots] u(\mathbf{p}) . \end{aligned} \quad (77)$$

between different baryons  $(B', B)$ . Here we have omitted contributions proportional to the momentum transfer  $q_\mu$ .

Knowledge of the  $g_A(B, B')$  and  $g_V(B, B')$  are crucial for the theory of baryon semi-leptonic decays. First let us consider the calculation of the axial vector matrix elements  $g_A(B, B')$ . Our main interest in this brief discussion will be to examine the effects of symmetry breaking. The leading order term (in  $1/N_C$ ) of the spatial components of the axial current is straightforwardly obtained to be

$$\int d^3r A_i^a = C D_{ai}(A) , \quad (78)$$

where  $A(t)$  is the collective coordinate matrix. The constant  $C$  denotes an integral over the chiral angle. We refer the interested reader to refs [57, 46] for the explicit expression. Then,

$$g_A^a(B', B) = C \langle B' | D_{a3} | B \rangle . \quad (79)$$

Table 3: The matrix elements of the axial-vector current (79) between different baryon states in the flavor symmetric limit. Displayed are both the strangeness conserving (a) and strangeness changing (b) processes. The first column gives the relevant flavor component of the axial current.

	(a)			
$A^{\pi^-}$	$n \rightarrow p$	$\Sigma^- \rightarrow \Lambda$	$\Sigma^- \rightarrow \Sigma^0$	$\Xi^- \rightarrow \Xi^0$
	$F + D$	$\frac{2}{\sqrt{6}}D$	$\sqrt{2}F$	$D - F$
	(b)			
$A^{K^-}$	$\Lambda \rightarrow p$	$\Sigma^- \rightarrow n$	$\Xi^- \rightarrow \Lambda$	$\Xi^- \rightarrow \Sigma^0$
	$\frac{1}{\sqrt{6}}(3F + D)$	$D - F$	$\frac{1}{\sqrt{6}}(3F - D)$	$\frac{1}{\sqrt{2}}(F + D)$

The flavor index  $a$  has to be chosen according to whether strangeness conserving ( $a = 1, 2, 3, 8$ ) or strangeness changing ( $a = 4, \dots, 7$ ) processes are being considered. The corresponding result for the axial charge of the nucleon  $g_A = g_A^{1+i2}(p, n)$ , as measured in neutron beta-decay, is predicted too low in many soliton models. This problem is already encountered in the two flavor model and gets worse in  $SU(3)$  as the Clebsch-Gordon coefficient associated with  $D_{1+i2\ 3}$  changes by a factor of 7/10. As symmetry breaking is increased the  $SU(3)$  prediction for  $g_A$  becomes larger [42]

$$g_A(SU(3)) = \frac{7}{10} [1 + 0.0514\gamma\beta^2 + \dots] g_A(SU(2)) . \quad (80)$$

Actually the exact treatment shows that with increasing symmetry breaking the two flavor result is approached, although only slowly. Taking everything together, including subleading terms in (79), finally gives  $g_A = 0.98$  for  $e = 4.0$  [46] which is about 4/5 of the experimental value  $g_A(\text{expt.}) = 1.26$ .

We may understand the tendency to approach the  $SU(2)$  limit for large  $\gamma\beta^2$  as follows. In the small  $SU(3)$  breaking case, there is just a small extra “cost” for producing an  $\bar{s}s$  pair rather than a  $\bar{u}u$  or  $\bar{d}d$  pair. As  $\gamma\beta^2$  gets larger it is more expensive to make an  $\bar{s}s$  pair and eventually  $\bar{s}s$  pairs should be absent from the nucleon wave-function, recovering the  $SU(2)$  picture.

Returning to the general case one should first note that flavor symmetry relates the octet axial current matrix elements between various baryons. Conventionally they are expressed using  $SU(3)$  covariance in terms of two unknown constants (or reduced matrix elements)  $F$  and  $D$ . One has to use models to determine these constants. In the flavor symmetric Skyrme model one finds [53]  $D/F = 9/5$  and  $D + F = 7C/15 = g_A$ . In table 3 the flavor symmetric dependences of the axial matrix elements on  $F$  and  $D$  are displayed. As one departs from the flavor symmetric case the baryon wave-functions acquire admixture from higher dimensional  $SU(3)$  representations making the  $SU(3)$  covariant parameterization in terms of  $F$  and  $D$  inadequate. In the presence of  $SU(3)$  symmetry breaking we must, without further assumptions, parameterize each decay amplitude separately. It is still reasonable to maintain the isospin invariance relations.

As an example of the perturbative corrections consider the axial  $\Lambda \rightarrow p$  transition in the Cabibbo scheme [56] for semi-leptonic hyperon decays. The analog of (59) for the  $\Lambda$  hyperon is

$$|\Lambda\rangle = |\Lambda, 8\rangle + \frac{3}{50}\gamma\beta^2|\Lambda, 27\rangle + \dots . \quad (81)$$

Noting that the  $D$ -functions mix different  $SU(3)$  representations, we get

$$\langle p \uparrow | D_{K-3} | \Lambda \uparrow \rangle = \frac{2}{5\sqrt{3}} - \frac{7\sqrt{3}}{1125}\gamma\beta^2 + \dots , \quad D_{K-3} = \frac{1}{\sqrt{2}}(D_{43} - iD_{53}) . \quad (82)$$

Of course, this expansion just provides a first approximation to the symmetry breaking dependence of the Cabibbo matrix elements. Using the exact treatment initiated by Yabu and Ando [44] this dependence can be computed numerically as shown in figure 2 for some processes of interest [57, 58]. Those results are normalized to the  $SU(3)$  symmetric values in table 3 to illustrate that the



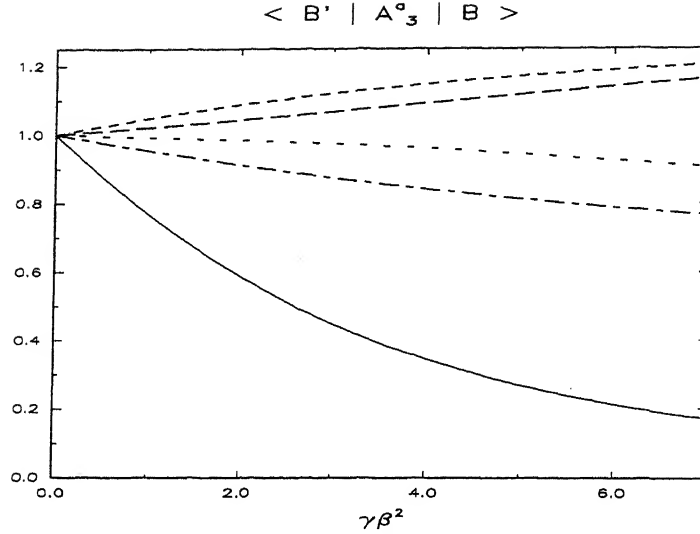


Figure 2: The variation of axial vector matrix elements with the effective symmetry breaking parameter  $\gamma\beta^2$ .

Full line:  $\langle p|\bar{s}\gamma_3\gamma_5 s|p\rangle$ ; dashed dotted line:  $\langle p|\bar{u}\gamma_3\gamma_5 s|\Lambda\rangle$ ; dotted line:  $\langle n|\bar{u}\gamma_3\gamma_5 s|\Sigma^-\rangle$ ; long dashed line:  $\langle \Lambda|\bar{u}\gamma_3\gamma_5 s|\Xi^-\rangle$ ; dashed line:  $\langle p|\bar{u}\gamma_3\gamma_5 d|n\rangle$ . These matrix elements, which are taken from refs [57] and [58], are normalized to the flavor symmetric values.

matrix elements vary in different ways with symmetry breaking. In this figure also the variation of the nucleon matrix element of the flavor conserving axial current  $H_3(0) = \langle N|\bar{s}\gamma_3\gamma_5 s|N\rangle$  is displayed. Obviously  $H_3(0)$  decreases very rapidly with increasing symmetry breaking. This is easily visualized, as mentioned above, as a reflection of the increased cost of making extra  $\bar{s}s$  pairs in the nucleon wave-function as  $\gamma\beta^2$  increases. On the contrary the Cabibbo matrix elements exhibit only a moderate dependence on  $\gamma\beta^2$ . It is this different behavior of the matrix elements that makes the application of exact flavor symmetry to the analyses of the EMC-SLAC-SMC experiments suspicious. Stated otherwise, the strange quark contribution to the nucleon matrix element of the axial singlet current (loosely “proton spin”) may be decreased significantly as a consequence of symmetry breaking without contradicting the successful Cabibbo scheme for the semi-leptonic decays of the hyperons.

As an interesting contrast to the axial matrix elements, consider the evaluation of the vector matrix elements  $g_V(B, B')$  needed for the hyperon semi-leptonic decays. The dominant contribution is given by the matrix elements of the  $SU(3)$  flavor generators

$$g_V^a(B', B) = \langle B'|L_a|B\rangle. \quad (83)$$

For example if we sandwich the generators  $L_{K^-}$  between the perturbative  $\Lambda$  and  $p$  states given in (81) and (59) and recognize that group generators can only connect states belonging to the same irreducible representation, we see that symmetry breaking corrections start out as  $(\gamma\beta^2)^2$  rather than  $\gamma\beta^2$ . This is just a demonstration of the Ademollo-Gatto theorem [59], which “protects” the vector matrix elements against small symmetry breaking corrections. Since  $\gamma\beta^2$  is large the numerical validity of this result is questionable. However, the exact Yabu-Ando scheme does confirm [57] that vector matrix elements suffer at most, 10% deviation from the symmetric values, even for large symmetry breaking, *e.g.*  $\gamma\beta^2 \approx 7$ .

A reduction of strangeness in the nucleon with increasing  $\gamma\beta^2$  is also predicted for the scalar strange content fraction of the proton

$$X_s = \frac{\langle p|\bar{s}s|p\rangle - \langle 0|\bar{s}s|0\rangle}{\langle p|\bar{u}u + \bar{d}d + \bar{s}s|p\rangle - \langle 0|\bar{u}u + \bar{d}d + \bar{s}s|0\rangle}. \quad (84)$$

Here the state  $|0\rangle$  refers to the soliton being absent. Models of quark flavor dynamics, as *e.g.* the one of Nambu-Jona-Lasinio [9], indicate that matrix elements of quark bilinears  $\bar{q}\lambda_a q$  may be

taken as proportional to the matrix elements of  $\text{tr} [\lambda_a (U + U^\dagger - 2)]$ . Then we straightforwardly get

$$X_s = \frac{1}{3} \langle p | 1 - D_{ss} | p \rangle \approx \frac{7}{30} - \frac{43}{2250} \gamma \beta^2 + \dots \quad (85)$$

In this case, however, the deviation from the flavor symmetric result [60] ( $X_s = 7/30$ ) is considerably mitigated [61] as compared to the variation of  $H_s$  defined in (76). The symmetry breaking has to be as large as  $\gamma \beta^2 \approx 4.5$  to obtain a reduction of the order of 50%. In the case of  $H_s$  this was already achieved for  $\gamma \beta^2 \approx 2.5$ . In any event, the additional quark-antiquark excitations in the nucleon, which are parametrized by the admixture of higher dimensional  $SU(3)$  representations (59), clearly tend to cancel the virtual strange quarks of the pure octet nucleon.

The three flavor Skyrme model under present consideration provides a convenient way to study the nucleon matrix elements of the vector current  $\bar{s} \gamma_\mu s$ . These are theoretically interesting because they would vanish in a pure valence quark model of the nucleon and so test finer details of nucleon structure. They are experimentally interesting because they can be extracted from measurements of the parity violating asymmetry in the elastic scattering of polarized electrons from the proton. The precise form factors needed are defined by

$$\langle P(p') | \bar{q}_s \gamma_\mu q_s | P(p) \rangle = \bar{u}(p') \left[ \gamma_\mu F_s(q^2) + \frac{\sigma_{\mu\nu} q^\nu}{2M_p} \tilde{F}_s(q^2) \right] u(p) \quad (86)$$

These form factors are currently under intensive experimental investigation, *cf.* refs [62] and have been estimated in various models. The models range from vector-meson-pole fits [63] of dispersion relations [64] through vector meson dominance approaches [46] and kaon-loop calculations with [65] and without [66] vector meson dominance contributions to soliton model calculations [46, 67, 68]. The numerical results for the strange magnetic moment  $\mu_S = \tilde{F}_s(0) \approx -0.31 \pm 0.09 \dots 0.25$  are quite diverse. The predictions for the strange charge radius  $r_S^2 = -6dF_s(q^2)/dq^2|_{q=0}$  are almost equally scattered  $r_S^2 \approx -0.20 \dots 0.14 \text{fm}^2$ .

In order to evaluate these form factors in the three flavor Skyrme model one requires the matrix elements of the “strange” combination

$$Q^s = \frac{1}{3} \mathbf{1} - \frac{1}{\sqrt{3}} \lambda_8 = Q^0 - \frac{2}{\sqrt{3}} Q^8 \quad (87)$$

rather than the electromagnetic one (67) between proton states. Using the same value  $e = 4.0$  as used consistently for the three flavor pseudoscalar model yields the predictions

$$\mu_S = -0.13 \text{n.m.}, \quad r_S^2 = -0.10 \text{fm}^2 \quad (88)$$

Here n.m. stands for nuclear magnetons. It should be stressed that these results are obtained within the Yabu-Ando approach, *i.e.* the proton wave-function contains sizable admixture of higher dimensional representations. If a pure octet wave-function were employed to compute the matrix elements of the collective operators the strange magnetic moment would have been  $\mu_S = -0.33$ . The proper inclusion of symmetry breaking into the nucleon wave-function is again seen to reduce the effect of the strange degrees of freedom in the nucleon. We already discussed above that as the strange quarks within the nucleon become more massive (effect of symmetry breaking) their excitation becomes less likely.

In the next few years a great deal of new experimental information on the form factors  $F_s$  and  $\tilde{F}_s$  should become available. This would enable more accurate comparison with (for a given effective meson Lagrangian) the essentially parameter free predictions of the soliton theory.

## 5 The nucleon as a vector meson soliton

According to the modern view of the Skyrme model approach we should start from the “full” effective Lagrangian which contains mesons of all spins. The practical criterion on which particles

to include is to find an effective Lagrangian which does a good job of explaining the low energy experimental data in the meson sector. On these grounds it is evident that the vector mesons should be included in the effective Lagrangian. In this section we give a brief sketch of the soliton sector of the Lagrangian of pseudoscalars and vectors (see section 3.3) and note that it leads to significant improvements of many predictions. In particular it is crucial for discussing the so-called *proton spin puzzle*.

### 5.1 Generalized soliton ansatz and profile functions

As a first step we construct the soliton of the Lagrangian defined in section 3.3. The generalization of the hedgehog *ansatz* (7) to the vector meson model requires the time component of the  $\omega$  field and the space components of the  $\rho$  field to be different from zero. Parity and grand spin symmetry<sup>9</sup> allow for three radial functions

$$\xi_\pi = \exp\left(\frac{i}{2}\hat{r} \cdot \tau F(r)\right), \quad \omega_0 = \frac{\omega(r)}{2g}, \quad \rho_i^a = \frac{G(r)}{gr} \epsilon_{ija} \hat{r}_j. \quad (89)$$

Substituting these *ansätze* into the action described in section 3.3 yields the classical mass,

$$\begin{aligned} E = 4\pi \int dr & \left[ \frac{f_\pi^2}{2} (F'^2 r^2 + 2\sin^2 F) - \frac{r^2}{2g^2} (\omega'^2 + m_\rho^2 \omega^2) + \frac{1}{g^2} [G'^2 + \frac{G^2}{2r^2} (G+2)^2] \right. \\ & + \frac{m_\rho^2}{g^2} (1+G-\cos F)^2 + \frac{\gamma_1}{g} F' \omega \sin^2 F - \frac{2\gamma_2}{g} G' \omega \sin F \\ & + \frac{\gamma_3}{g} F' \omega G (G+2) + \frac{1}{g} (\gamma_2 + \gamma_3) F' \omega [1 - 2(G+1)\cos F + \cos^2 F] \\ & + (1 - \cos F) \left\{ 4\delta' r^2 + 2(2\beta' - \frac{\alpha'}{g^2}) (F'^2 r^2 + 2\sin^2 F) \right. \\ & \left. \left. - \frac{2\alpha'}{g^2} [\omega^2 r^2 - 2(G+1-\cos F)^2 - 4(1+\cos F)(1+G-\cos F)] \right\} \right]. \quad (90) \end{aligned}$$

Application of the variational principle to this functional yields second order coupled non-linear differential equations for the radial functions  $F(r)$ ,  $\omega(r)$  and  $G(r)$ . The boundary conditions for the chiral angle  $F(r) = \pi$  and  $F(\infty) = 0$ , which correspond to unit baryon number, also determine the boundary conditions of the vector meson profiles via the differential equations and the requirement of finite energy. For example we find  $G(0) = -2$ . A typical set of resulting profile functions is shown in figure 3.

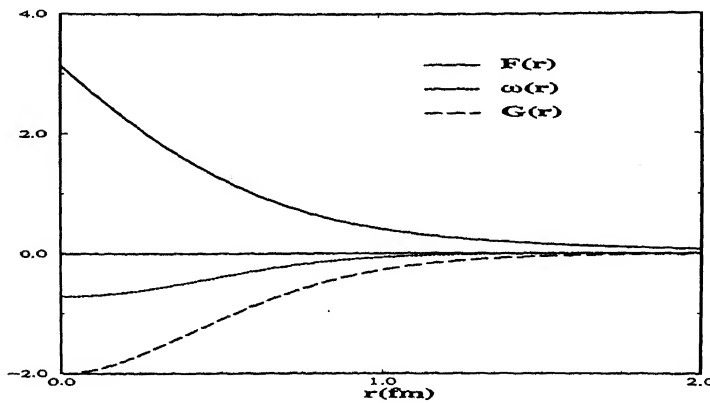


Figure 3: The profile functions which minimize the classical vector meson energy functional (90) for the parameters  $g = 5.85$ ,  $\tilde{h} = 0.4$ ,  $\tilde{g}_{VV\pi} = 1.9$  and  $\kappa = 1.0$ .  $\omega(r)$  is measured in units of  $m_V = \sqrt{2}gf_\pi$ .

<sup>9</sup>The grand spin  $J + I$  is the characteristic invariance of the hedgehog ansatz.

As in the pseudoscalar model we have to generate states with good spin and isospin from this classical field configuration. To start with, one introduces collective coordinates (9) for all fields which have non-vanishing spin or isospin. However, an additional complication arises because there are vector meson field components which vanish classically but get excited by the collective rotation. In the two flavor case the appropriate *ansatz* for these excitations reads

$$\rho_0 = \frac{1}{2g} A(t) [\xi_1(r) \Omega + \xi_2(r) (\hat{r} \cdot \Omega) \hat{r}] A^\dagger(t), \quad \omega_i = \frac{\Phi(r)}{2g} \epsilon_{ijk} \Omega_j \hat{r}_k, \quad (91)$$

where the angular velocity of the rotating soliton,  $\Omega_i$  is defined in (11). The three radial functions  $\xi_1, \xi_2$  and  $\Phi$  are not the only ones which get excited. As these radial functions are non-zero they provide sources for the non-strange component of the iso-singlet pseudoscalar field via the  $\epsilon$ -terms (43). In the two flavor formulation the appropriate *ansatz* which takes into account the pseudoscalar nature of the  $\eta$  field reads

$$U(r, t) = e^{i\eta_T(r)} A(t) U_0(r) A^\dagger(t) \quad \text{with} \quad \eta_T(r) = \frac{1}{f_\pi} \eta(r) \hat{r} \cdot \Omega. \quad (92)$$

As we will observe shortly, the excitation of this  $\eta$  field plays a decisive role in the context of the *proton spin puzzle*. The additional radial functions are determined from extremizing the moment of inertia for rotations in coordinate space,

$$\begin{aligned} \alpha^2 = & \frac{8\pi}{3} \int dr \left\{ f_\pi^2 r^2 \sin^2 F - \frac{4}{g^2} (\phi'^2 + 2 \frac{\phi^2}{r^2} + m_\rho^2 \phi^2) + \frac{m_\rho^2}{2g^2} r^2 [(\xi_1 + \xi_2)^2 + 2(\xi_1 - 1 + \cos F)^2] \right. \\ & + \frac{1}{2g^2} [(3\xi_1'^2 + 2\xi_1' \xi_2' + \xi_2'^2) r^2 + 2(G^2 + 2G + 2)\xi_2^2 + 4G^2(\xi_1^2 + \xi_1 \xi_2 - 2\xi_1 - \xi_2 + 1)] \\ & + \frac{4}{g} \gamma_1 \phi F' \sin^2 F + \frac{4}{g} \gamma_3 \phi F' [(G - \xi_1)(1 - \cos F) + (1 - \cos F)^2 - G\xi_1] \\ & + \frac{2\gamma_2}{g} \left\{ \phi' \sin F (G - \xi_1 + 2 - 2\cos F) + \phi \sin F (\xi_1' - G') \right. \\ & \left. + \phi F' [2 + 2\sin^2 F + (\xi_1 - G - 2)\cos F - 2(\xi_1 + \xi_2)] \right\} \\ & - \frac{1}{2} [\eta'^2 r^2 + 2\eta^2 + m_\eta^2 r^2 \eta^2] + \frac{\gamma_2 g}{2f_\pi} [\eta(\phi\omega' - \omega\phi') - \eta' \phi\omega] \\ & - \frac{\gamma_1}{3gf_\pi} [\eta'(\xi_1 + \xi_2) \sin^2 F + 2\eta F' (G + \xi_1) \sin F] - \frac{3\gamma_3}{gf_\pi} \eta' (G + 1 - \cos F)^2 (\xi_1 + \xi_2) \\ & \left. - \frac{\gamma_2}{gf_\pi} \left\{ \eta' [(G + \xi_1)G + (\xi_1 + \xi_2)[(1 - \cos F)^2 - 2G\cos F]] + \eta(G\xi_1' - G'\xi_1) \right\} \right\}, \quad (93) \end{aligned}$$

together with suitable boundary conditions. In eq (93) we have not displayed the explicit contributions from the symmetry breakers (which are in fact small). We will mostly limit the present discussion to the two flavor case. In the case of three flavor vector meson models the situation is even more complicated as also  $K^*$  type fields get excited. Also there will be additional symmetry breakers on the level of the collective Lagrangian which are of the form  $\sum_{i=1}^3 D_{8i} \Omega_i$  and stem from terms which are linear in the time derivative. They can straightforwardly be implemented in the collective quantization approach. Here we will omit details but rather refer the reader to the literature [67, 40]. The general pattern for computing baryon properties is essentially the same as that discussed for the Lagrangian of only pseudoscalars in section 4.

## 5.2 Axial singlet current and proton spin puzzle

Notice that in (93) we included by hand a mass term for the rotationally excited profile  $\eta(r)$  of a pseudoscalar isosinglet field. Actually the existence of such a term has not yet been justified. Before proceeding we must do so since the term turns out to be very important.

In section 3.1 we mentioned that the QCD axial singlet current

$$J_{5,\mu}^{(0)} = \bar{u}\gamma_\mu\gamma_5 u + \bar{d}\gamma_\mu\gamma_5 d + \bar{s}\gamma_\mu\gamma_5 s, \quad (94)$$

is not conserved even for zero quark masses:  $\partial^\mu J_{5,\mu}^{(0)} = G$ , where the  $U_A(1)$  anomaly  $G$  is proportional to the product of the QCD field strength tensor  $F_{\mu\nu}^a$  and its dual. In order to mock up this non-conservation equation at the effective Lagrangian level [69] we may add the terms

$$\frac{c}{2}G^2 + \frac{iG}{12}\ln\left(\frac{\det U}{\det U^\dagger}\right), \quad (95)$$

where  $G$  is now considered a composite glueball field which “dominates” the  $U_A(1)$  anomaly. Here we assumed three light flavors and also that the strong CP violation parameter  $\theta$  is zero. Furthermore it is necessary that, except for the terms representing quark mass symmetry breaking, all the other terms in the effective Lagrangian be invariant under  $U_A(1)$ . The parameter  $c$  above is determined by

$$m_{\eta_0}^2 \approx \frac{1}{6cf_\pi^2}, \quad (96)$$

in the approximation where the quark mass terms are neglected.  $\eta_0$  is the  $SU(3)$  singlet pseudoscalar field as in (45). This equation arises after noting that  $G$  is like an auxiliary field and may be integrated out:  $G = \eta_0/(\sqrt{6}cf_\pi)$ . In the effective Lagrangian the realization of the axial singlet current, obtained by a Noether variation, is

$$J_{5,\mu}^{(0)} = \sqrt{6}f_\pi\partial_\mu\eta_0 + \tilde{J}_{5,\mu}^{(0)}. \quad (97)$$

Here the first term is the contribution from the pseudoscalar field and the second term is due to the addition of vector fields.  $\tilde{J}_{5,\mu}^{(0)}$  has a complicated structure but, in particular, contains *non-derivative* terms like  $\epsilon^{\mu\nu\alpha\beta}\text{tr}(\rho_\nu\rho_\alpha\rho_\beta)$ . Using this decomposition we may write the equation of motion for the  $\eta_0$  field as

$$(\partial^2 + m_{\eta_0}^2)\eta_0 = \frac{1}{\sqrt{6}f_\pi}\partial^\mu\tilde{J}_{5,\mu}^{(0)}, \quad (98)$$

which shows that the vector meson contribution to the axial singlet current may act as a source for a non-trivial excitation associated with the  $\eta_0$  field in the soliton sector.

Now the form factors for the proton matrix elements of the axial singlet current are obviously just the sums of the three separate form factors introduced in (76):

$$H(q^2) = H_u(q^2) + H_d(q^2) + H_s(q^2) \quad \text{and} \quad \tilde{H}(q^2) = \tilde{H}_u(q^2) + \tilde{H}_d(q^2) + \tilde{H}_s(q^2). \quad (99)$$

If the vector mesons are not present, eq (97) shows that the operator for the axial singlet current must be (even in the soliton sector) a pure derivative. This means that, regardless of the details of the calculation, the matrix element for the sum of the three terms in (76) must be proportional to the momentum transfer  $q_\mu$ . Thus  $\tilde{H}(q^2)$  is non-zero and  $H(q^2) = 0$ . From the theory of Dirac particles we recognize that the quantity  $H(0)$  has the interpretation of twice the quark spin part of the proton's angular momentum. We see that the Skyrme model of only pseudoscalars predicts that the expectation value of the net quark spin operator vanishes; the total angular momentum (1/2) of the proton must involve, at a fundamental level, the rotational and gluonic pieces! Note that the above argument for  $H(0) = 0$  with the Lagrangian of only pseudoscalars continues to hold even if symmetry breaking contributions are taken into account [58].

The situation is a little different when vector mesons are included in the effective Lagrangian. Since  $\tilde{J}_{5,\mu}^{(0)}$  has pieces which are not pure derivatives it then is possible to obtain  $H(0) \neq 0$ . A convenient parameterization for this calculation in the effective Lagrangian model is

$$\langle P(\mathbf{p}') | \tilde{J}_{5,i}^{(0)} | P(\mathbf{p}) \rangle = H(q^2) \langle 2J_i \rangle. \quad (100)$$

Once all the radial functions have been determined as before from the appropriate variational principles, it is straightforward to compute  $H(0)$  from eq (100). One only has to recall that under the collective coordinate quantization the angular momentum operator is given by  $\mathbf{J} = \mathbf{\Omega}/\alpha^2$ . The numerical results for a variety of allowed parameters (*cf.* discussion after eq (43)) are displayed in table 4. Surprisingly the predictions of the vector meson model for  $H(0)$  are very robust against

Table 4: Predictions for the matrix element of the proton axial singlet current for various allowed sets of parameters in the vector meson model. Results are taken from ref [70].

$\tilde{h}$	0.4	0.4	0.4	0.7	0.5	0.2	0.1
$\tilde{g}_{VV\Phi}$	1.9	1.9	1.9	2.2	2.0	1.7	1.5
$\kappa$	0.0	0.5	1.0	0.0	0.0	0.0	0.0
$H(0)$	0.34	0.33	0.30	0.29	0.34	0.32	0.28

possible changes of the parameters of the model.

Even though we get a non-zero value for  $H(0)$  in the vector meson model it is still small compared to  $H(0) = 1$ , the expectation from the simple non-relativistic quark model. Qualitatively the soliton model results with and without vectors are similar. Since one has a natural prejudice that the quark model results should be roughly correct, this would at first seem to be a serious defect of the soliton approach to nucleon properties.

Of course one can only make an accurate judgment on the matter by appealing to experiment.  $H(0)$  can be found from eq (99) if we can experimentally obtain separately  $H_u(0)$ ,  $H_d(0)$  and  $H_s(0)$ . The linear combination

$$H_u(0) - H_d(0) = g_A = 1.257, \quad (101)$$

is reliably obtained by an isotopic spin rotation of the axial form factor describing neutron beta-decay. Similarly the estimate for the “eighth” octet component

$$H_u(0) + H_d(0) - 2H_s(0) \approx 0.575 \pm 0.016, \quad (102)$$

may be gotten from an  $SU(3)$  flavor rotation of the data on hyperon beta-decay experiments. Clearly one more linear combination is needed in order to disentangle the individual  $H_i(0)$  and that situation existed for many years. About ten years ago different experimental groups (EMC, SLAC, SMC) used polarized lepton beams to probe the structure of nucleons. The deep inelastic scattering data [55] were used to extract the combination

$$4H_u(0) + H_d(0) + H_s(0), \quad (103)$$

in which the axial current form factor for each quark is weighted proportionally to the square of the quark electric charge. Combining these data resulted in  $H(0) \approx 0.3$  [71]. The experimental results were later on confirmed [72, 73, 74]. However, it turns out that the theoretical extraction of  $H(0)$  is quite complicated as it involves a careful treatment of perturbative QCD corrections. The value [75]

$$H(0) = 0.27 \pm 0.04 \quad (104)$$

is nowadays considered correct. At the time this low value was considered hard to understand and the situation was called the *proton spin puzzle*. We have just seen that the soliton approach does however provide a simple explanation of such a low value.

Clearly the prediction of the vector meson treatment described in Table 5.1, yielding  $H(0)$  about 0.30, is in good agreement with the data. From this we learn two things. First, the simplest quark model does not give a good description of the spin structure of the nucleon. Second, the soliton approach based on an effective Lagrangian including vector mesons markedly improves the qualitatively reasonable predictions of the soliton treatment based on a pseudoscalars only

effective Lagrangian. A physical interpretation of the latter statement is that the pseudoscalars only Lagrangian mainly probes the “pion cloud” of the nucleon while the vector Lagrangian probes a little more deeply.

For completeness we remark on a possible caveat. The estimate of (102) is based on the use of exact  $SU(3)$  symmetry. However in Fig 4.1 of section 4.4 we showed that precisely this current matrix element is expected to exhibit stronger suppression than others due to  $SU(3)$  symmetry breaking. Nevertheless it turns out that [70] the numerical evaluation of  $H(0)$  is not very sensitive to this feature. This is to be contrasted with the behavior of  $H_s(0)$ , which decreases rapidly with symmetry breaking, *cf.* section 4.4.

### 5.3 Other improvements with vector mesons

The famous problem of explaining the neutron-proton mass difference is another one which requires the addition of vector mesons to the effective Lagrangian in order to obtain a satisfactory solution in the nucleon-as-soliton picture. It is known that the electromagnetic (*i.e.* one photon loop) contribution has the wrong sign. After correcting for the electromagnetic interaction the remaining “strong” part of the neutron-proton mass-difference should be  $(M_n - M_p)_{\text{strong}} \approx (2.0 \pm 0.3)\text{MeV}$  [76]. At the quark level this arises from the down quark-up quark mass difference, controlled by the parameter  $y$  in eqs (20) and (21). Information on  $y$  can be most easily gained by analyzing the  $K^+ - K^0$  mass-difference, yielding  $y \approx (-0.4 \dots -0.2)$  [33]. To understand the problem it is helpful to consider the contribution of the (presumably dominating)  $\delta'$ -type symmetry breaker to the neutron-proton mass-difference. Since the  $d$ - $u$  quark mass difference clearly exists with only two flavors it is interesting to first consider the problem at this level. Then the relevant piece of the  $\delta'$  term is proportional to

$$\text{tr} [\tau_3 (U + U^\dagger)] . \quad (105)$$

Using the ansatz (92) we see that  $U = \exp(i\eta_T)[\cos(\psi) + i\mathbf{n} \cdot \boldsymbol{\tau}\sin(\psi)]$ , where  $\psi$  is some angle. Then (105) is proportional to  $\sin(\eta_T)$ . In other words the contribution vanishes unless the field  $\eta_T$  gets excited due to the collective rotation (or any kind of symmetry breaking). Now (98) together with (97) shows that this will not happen if only pseudoscalars are present in the effective Lagrangian; the vector meson contribution  $\tilde{J}_{5,\mu}^{(0)}$  must also be present. This is analogous to the situation concerning the *proton spin puzzle*. The contribution of the  $\delta'$  term turns out to be

$$(M_n - M_p)_{\text{strong}} = -\frac{2y\delta'}{3\alpha^2} \int d^3r \sin F(r) \eta(r) + \dots . \quad (106)$$

Using the full two-flavor vector meson result for  $\tilde{J}_{5,\mu}^{(0)}$  which was already employed to compute  $H(0)$  yields [77]

$$(M_n - M_p)_{\text{strong}} \approx 1.4\text{MeV} \quad (107)$$

which, not surprisingly, turns out to be about as robust against changes of the parameters as is  $H(0)$ . This prediction is still somewhat too small when compared to the empirical value. However, it turns out that the missing  $\sim 0.5\text{MeV}$  can be attributed to three flavor effects as matrix elements of  $D_{38}$  are non-vanishing<sup>10</sup>.

The addition of vector mesons also plays an important role in the discussion of the “sizes” of the nucleons: the nucleon radii. As can be observed from table 2 the Skyrme model of pseudoscalars only seriously underestimates the empirical values for the baryon radii. The presence of the  $\omega$  meson provides an increase of the isoscalar radius [52]

$$\langle r^2 \rangle_{I=0} \approx \langle r^2 \rangle_B + \frac{6}{m_\rho^2} , \quad (108)$$

<sup>10</sup>It should be remarked that  $\langle p|D_{38}|p \rangle$  quickly approaches zero as  $SU(3)$  symmetry breaking is increased. This decreases  $SU(3)$  type contributions to (107).

where  $\langle r^2 \rangle_B$  is the radius associated with baryon number current (3). The additional piece in eq (108) is a consequence of (approximate) vector meson dominance in this model [78], which indeed is observed when including the vector mesons in a chirally invariant manner. As can be seen from table 2 this increase of about  $0.35\text{fm}^2$  will significantly improve the predictions for the radii.

A similar interesting improvement due to vector mesons is obtained in the context of meson-baryon scattering. In these investigations one introduces small fluctuations off the classical soliton. Eventually these fluctuations are quantized to represent in- and out-going meson fields, thereby determining the scattering matrix [79]. It turns out that in the pseudoscalar Skyrme model the phase-shifts extracted from this scattering matrix rise almost linearly with the momentum of the in-going pion. This undesired feature is mostly due to the contact interaction between pions contained in the Skyrme model Lagrangian (*cf.* section 2). When introducing vector mesons this contact interaction is essentially replaced by the exchange of such a vector meson,

$$\frac{-1}{m_\rho^2} \longrightarrow \frac{1}{q^2 - m_\rho^2}. \quad (109)$$

As this interaction decreases for large momentum transfers,  $q^2$  the resulting phase-shifts assume a constant value for large energies rather than rising linearly [80]. Clearly this effect is similar to the one observed when going from the Fermi to the standard model of electro-weak interactions.

These examples show that while the inclusion of vector meson degrees of freedom involves quite a few technical details it clearly provides a more realistic picture of the nucleon as a chiral soliton.

## 6 Summary and discussion

Aside from the mass spectra and current matrix elements of the low-lying  $\frac{1}{2}^+$  and  $\frac{3}{2}^+$  baryons treated here the soliton approach has been extensively employed to study meson nucleon scattering [79, 80], baryons containing a heavy quark [81], nucleon-nucleon scattering [82], few nucleon systems [83] and nuclear matter [84].

In the present survey, we started out with a historical introduction (section 1) and a concise technical summary of the original two flavor Skyrme model (section 2). In these sections the physical interpretation and justification of the model were emphasized; it is hoped that they will be useful to beginners in this area of research (see also [85] and [86]).

We next attempted to develop the generalization of the original Skyrme model which is suggested by the large  $N_C$  approximation to QCD. In this approach the Skyrme Lagrangian is to be replaced by a more general effective Lagrangian containing mesons of all spins. Perhaps some day an analytic expression in this framework will be found. At present it seems necessary to obtain an approximation based on including the lowest energy resonances and constraining the model by the symmetries of the underlying QCD. The concept of chiral symmetry which plays a crucial role in this extension was explored in section 3. Furthermore the original Skyrme model of two light flavors was extended to three flavors, as it is now well established that the nucleons belong to a flavor  $SU(3)$  multiplet.

It is worthwhile to stress that once the effective Lagrangian has been determined from the meson sector, the soliton approach provides in principle a *zero parameter* description of baryon properties (In our case we introduced just one parameter which had to be fit from the baryon sector.).

In section 4 we studied the technical tools needed to treat the flavor  $SU(3)$  symmetry and its breaking at the (collective) baryon level. These were applied to the calculation of various interesting baryon matrix elements. Finally section 5 sketched the treatment of baryons based on an effective Lagrangian which also included the vector mesons. An application to the so-called *proton spin puzzle* demonstrated that the soliton approach seems to give a neat description of, otherwise hard to explain, experimental results on the quark spin structure of the nucleon. The improvements one encounters on including the vector mesons are in accord with the intuitive notion that the addition of higher mass resonances in the meson sector leads to a progressively more detailed understanding of the short distance structure of the nucleon-as-soliton.



We are happy to acknowledge the stimulating interactions we have had with many collaborators and colleagues while doing research related to the topics reviewed here.

## References

- [1] The early paper in the series which Skyrme himself evidently believed to be the key one is T. H. R. Skyrme, Proc. Roy. Soc. (London) **A260**, 127 (1961). Some additional properties of the nucleons were discussed in T. H. R. Skyrme Nucl. Phys. **31**, 556 (1961). A late paper in the series is T. H. R. Skyrme, J. Math. Phys. **12**, 1735 (1971).
- [2] A nice pedagogical treatment is provided by the text book R. Rajaraman, *Solitons and Instantons*, (1982), North-Holland.
- [3] J. G. Williams, J. Math. Phys. **11**, 2611 (1970).
- [4] N. K. Pak and H. C. Tze, Ann. Phys. (N.Y.) **117**, 164 (1979).
- [5] A. P. Balachandran, V. P. Nair, S. G. Rajeev and A. Stern, Phys. Rev. D **27**, 1153 (1983).
- [6] G. S. Adkins, C. R. Nappi and E. Witten, Nucl. Phys. **B228**, 552 (1983).
- [7] M. Gell-Mann and Y. Ne'eman, *The Eightfold Way*, Benjamin, New York, 1964.
- [8] E. C. G. Sudarshan and R. E. Marshak, Proceedings of the Padua conference on mesons and recently discovered particles, p. V-14 (1957). See also R. P. Feynman and M. Gell-Mann, Phys. Rev. **109**, 193 (1958) and J. J. Sakurai, Nuovo Cimento **7**, 649 (1958).
- [9] Y. Nambu and G. Jona-Lasinio, Phys. Rev. **122**, 345 (1961); **124**, 246 (1961).
- [10] M. Gell-Mann and M. Levy, Nuovo Cimento **16**, 705 (1960). See also K. Nishijima, Nuovo Cimento **11**, 698(1959) and F. Gursey, Nuovo Cimento **16**, 230 (1960).
- [11] J. Cronin, Phys. Rev. **161**, 1483 (1967); S. Weinberg, Phys. Rev. Lett. **18**, 188 (1967).
- [12] S. Weinberg, Physica **A96**, 327 (1979); J. Gasser and H. Leutwyler, Ann. Phys. (N.Y.) **158**, 142 (1984), Nucl. Phys. **B250**, 465 (1985).
- [13] Older work on chiral dynamics with baryons is summarized in S. Gasiorowicz and D.A. Geffen, Rev. Mod. Phys. **41**, 531 (1969). A recent revival is summarized in V. Bernard, N. Kaiser, Ulf-G. Meißner, Int. J. Mod. Phys. **E4**, 193 (1995).
- [14] G. 't Hooft, Nucl. Phys. **B72**, 461 (1974); **B75**, 461 (1975).
- [15] E. Witten, Nucl. Phys. **B160**, 57 (1979).
- [16] T. H. R. Skyrme, Int. J. Mod. Phys. **A3**, 2745 (1988). This talk was reconstructed by I. Aitchison.
- [17] R. H. Dalitz, Int. J. Mod. Phys. **A3**, 2719 (1988).
- [18] G. H. Derrick, J. Math. Phys. **5**, 1252 (1964).
- [19] This is treated in the book W. Pauli, *Meson Theory of Nuclear Forces*, Interscience Publishers, Inc., New York, 1946.
- [20] B. Moussallam and D. Kalafatis, Phys. Lett. **B272**, 196 (1991); G. Holzwarth, Phys. Lett. **B291**, 218 (1992); B. Moussallam, Ann. Phys. (NY) **225**, 264 (1993); G. Holzwarth, Nucl. Phys. **A572**, 69 (1994); H. Weigel, R. Alkofer and H. Reinhardt, Nucl. Phys. **A582**, 484 (1995); F. Meier and H. Walliser, Phys. Rep. **289**, 383 (1997).

- [21] R. Dashen, E. Jenkins and A. V. Manohar, Phys. Rev. **D49**, 4713 (1994).
- [22] N. Dorey, J. Hughes and M. Mattis, Phys. Rev. **D50**, 5816 (1994).
- [23] M. Bander and F. Hayot, Phys. Rev. **D30**, 1837 (1984); E. Braaten and J. P. Ralston, Phys. Rev. **D31**, 598 (1985).
- [24] C. Caso *et al.*, (Particle Data Group), Eur. Phys. J. **C3**, 1 (1998).
- [25] For a review see M. Neubert, Phys Rep. **245**, 259 (1994).
- [26] M. Gell-Mann, Phys. Rev. **125**, 1067 (1962); M. Gell-Mann, R. Oakes and B. Renner, *ibid* **175**, 2195 (1968).
- [27] S. L. Adler, Phys. Rev. **177** (1969) 2426; J. S. Bell and R. Jackiw, Nuov. Cim. **60A** (1969) 47.
- [28] W. Bardeen, Phys. Rev. **184**, 1848 (1969); B. Zumino, Wu Yong-Shi and A. Zee, Univ. of Washington preprint 4048-18-P3 (May 1983).
- [29] D. Ebert and H. Reinhardt, Nucl. Phys. **B271**, 188 (1986).
- [30] E. Witten, Nucl. Phys. **B223** (1983) 422, 433.
- [31] J. Wess and B. Zumino, Phys. Lett. **37B**, 95 (1971).
- [32] Ö. Kaymakçalan, S. Rajeev and J. Schechter, Phys. Rev. **D30**, 594 (1984); Ö. Kaymakçalan and J. Schechter Phys. Rev. **D31**, 1109 (1985).
- [33] J. Schechter, A. Subbaraman and H. Weigel, Phys. Rev. **D48**, 339 (1993); M. Harada and J. Schechter, Phys. Rev. **D54**, 3394 (1996).
- [34] S. Callan, S. Coleman, J. Wess and B. Zumino, Phys. Rev. **177**, 2247 (1969).
- [35] P. Jain, R. Johnson, Ulf-G. Meißner, N. W. Park and J. Schechter, Phys. Rev. **D37**, 3252 (1988).
- [36] Ulf-G. Meißner, N. Kaiser, H. Weigel and J. Schechter, Phys. Rev. **D39**, 1956 (1989).
- [37] M. Bando, T. Kugo and K. Yamawaki, Phys. Rep. **64**, 217 (1988).
- [38] J. Carlson, Nucl. Phys. **B253**, 149 (1985); **B277**, 253 (1986); P. Jain, J. Schechter and R. Sorkin, Phys. Rev. **D39**, 998 (1989); **D41**, 3855 (1990).
- [39] A. P. Balachandran, F. Lizzi, V. Rodgers and A. Stern, Nucl. Phys. **B256**, 525 (1985).
- [40] H. Weigel, Int. J. Mod. Phys. **A11**, 2419 (1996).
- [41] A. V. Manohar, Nucl. Phys. **B248**, 19 (1984).
- [42] N. W. Park, J. Schechter and H. Weigel, Phys. Lett. **B224**, 171 (1989).
- [43] J. de Swart, Rev. Mod. Phys. **35**, 916 (1963).
- [44] H. Yabu and K. Ando, Nucl. Phys. **B301**, 601 (1988).
- [45] H. Weigel, J. Schechter, N. W. Park and Ulf-G. Meißner, Phys. Rev. **D42**, 3177 (1990).
- [46] N. W. Park, J. Schechter and H. Weigel, Phys. Rev. **D43**, 869 (1991).
- [47] S. Okubo, Prog. Theor. Phys. **27**, 949 (1962).
- [48] E. Guadagnini, Nucl. Phys. **B236**, 15 (1984).

- [49] M. Prasałowicz, Phys. Lett. **158B**, 264 (1983).
- [50] M. Chemtob, Nucl. Phys. **B256**, 600 (1985).
- [51] E. Braaten, S.-M. Tse and C. Willcox, Phys. Rev. **D34**, 1482 (1986).
- [52] Ulf-G. Meißner, N. Kaiser and W. Weise, Nucl. Phys. **A466**, 685 (1987); Ulf-G. Meißner, Phys. Rep. **161**, 213 (1988).
- [53] G. S. Adkins and C. R. Nappi, Nucl. Phys. **B249**, 507 (1985).
- [54] B. Schwesinger and H. Weigel, Nucl. Phys. **A540**, 461 (1992).
- [55] J. Ashman *et al.*, Phys. Lett. **B206**, 364 (1988), Nucl. Phys. **B328**, 1 (1989).
- [56] N. Cabibbo, Phys. Rev. Lett. **10**, 531 (1963).
- [57] N. W. Park, J. Schechter and H. Weigel, Phys. Rev. **D41**, 2836 (1990).
- [58] N. W. Park, J. Schechter and H. Weigel, Phys. Lett. **B228**, 420 (1989).
- [59] M. Ademollo and R. Gatto, Phys. Rev. Lett. **13**, 264 (1964).
- [60] J. Donoghue and C. R. Nappi, Phys. Lett. **168B**, 105 (1986).
- [61] H. Yabu, Phys. Lett. **B218**, 124 (1989).
- [62] B. Müller *et al.* Phys. Rev. Lett. **78**, 3824 (1997); K. Aniol *et al.* Phys. Rev. Lett. **82**, 1096 (1999).
- [63] R. L. Jaffe, Phys. Lett. **B229**, 275 (1989).
- [64] G. Höhler *et al.*, Nucl. Phys. **B114**, 505 (1974).
- [65] M. J. Musolf and T. W. Donnelly, Z. Phys. **C57**, 559 (1993).
- [66] H. Forkel, M. Nielsen, X.-M. Jin and T. Cohen, Phys. Rev. **C50**, 3108 (1994).
- [67] N. W. Park and H. Weigel, Nucl. Phys. **A541**, 453 (1992).
- [68] H. Weigel, A. Abada, R. Alkofer and H. Reinhardt, Phys. Lett. **B353**, 20 (1995).
- [69] C. Rosenzweig, J. Schechter and G. Trahern, Phys. Rev. **D21**, 3388 (1980); P. Di Vecchia and G. Veneziano, Nucl. Phys. **B171**, 253 (1980); P. Nath and R. Arnowitt, Phys. Rev. **D23**, 1789 (1981); E. Witten, Nucl. Phys. **B156**, 269 (1979); A. Aurilia, Y. Takahashi and D. Townsend, Phys. Lett. **95B**, 265 (1980); K. Kawarabayashi and N. Ohta, Nucl. Phys. **B175**, 477 (1980).
- [70] R. Johnson, N. W. Park, J. Schechter, V. Soni and H. Weigel, Phys. Rev. **D42**, 2998 (1990).
- [71] S. Brodsky, J. Ellis and M. Karliner, Phys. Lett. **B206**, 309 (1988); J. Ellis and M. Karliner, Phys. Lett. **B213**, 73 (1988).
- [72] D. Adams *et al.*, Phys. Rev. **D56** (1997) 5330. B. Adeva *et al.*, Phys. Lett. **B412** (1997) 414.
- [73] K. Abe *et al.*, Phys. Rev. Lett. **76** (1996) 587.
- [74] K. Abe *et al.* Phys. Rev. **D58** (1998) 112003.
- [75] J. Ellis and M. Karliner, *The Strange Spin of the Nucleon*, hep-ph/9601280.
- [76] J. Gasser and H. Leutwyler, Ann. Phys. (NY) **158**, 142 (1984).
- [77] P. Jain, R. Johnson, N. W. Park, J. Schechter and H. Weigel, Phys. Rev. **D40**, 855 (1989).

- [78] J. Schechter, Phys. Rev. **D34**, 868 (1986).
- [79] H. Walliser and G. Eckart, Nucl. Phys. **A429**, 514 (1984); A. Hayashi, G. Eckart, G. Holzwarth and H. Walliser, Phys. Lett. **B147**, 5 (1984); M. P. Mattis and M. Karliner, Phys. Rev. **D31**, 2833 (1985).
- [80] B. Schwesinger and H. Weigel, Nucl. Phys. **A465**, 733 (1987); B. Schwesinger, H. Weigel, G. Holzwarth and A. Hayashi, Phys. Rep. **173**, 173 (1989);
- [81] C. G. Callan and I. Klebanov, Nucl. Phys. **B262**, 365 (1985); A compilation of references on heavy quark solitons may be found in: M. Harada, F. Sannino, J. Schechter and H. Weigel, Phys. Rev. **D56**, 4098 (1997).
- [82] A. Jackson, A. D. Jackson, A. S. Goldhaber, G. E. Brown, and L. C. Castillejo, Phys. Lett. **154B**, 101 (1985); A. Jackson, A. D. Jackson and V. Pasquier, Nucl. Phys. **A432**, 567 (1985); R. Vinh Mau, M. Lacombe, B. Loiseau, W. N. Cottingham and P. Lisboa, Phys. Lett. **150B**, 259 (1985); H. Yabu and K. Ando, Prog. Theor. Phys. **74**, 750 (1985) 750; H. Yabu, B. Schwesinger and G. Holzwarth, Phys. Lett. **B224**, 25 (1989).
- [83] H. Weigel, B. Schwesinger and G. Holzwarth, Phys. Lett. **B168**, 321 (1986); V. B. Kopeliovich and B. E. Stern, JETP Lett. **45**, 203 (1987); E. Braaten, L. Carson and S. Townsend, Phys. Lett. **B235**, 147 (1990); W. Y. Crutchfield, N. J. Snyderman, V. R. Brown, Phys. Rev. Lett. **68**, 1660 (1992); T. Wainzdoch and J. Wambach, Nucl. Phys. **A602**, 347 (1996); N. Walet, Nucl. Phys. **A606**, 429 (1996).
- [84] A. D. Jackson and J. J. M. Verbaarschot, Nucl. Phys. **A484**, 419 (1988); H. Forkel *et al.*, Nucl. Phys. **A504**, 818 (1989).
- [85] G. Holzwarth and B. Schwesinger, Rep. Prog. Phys. **49**, 825 (1986).
- [86] I. Zahed and G. E. Brown, Phys. Rep. **142**, 481 (1986).



## Part C : Formal Methods In QFT

14. Euclidean Methods In Quantum Field Theory by R.Ramanathan
15. Topics In Finite Temperature Field Theory by Ashoke Das
16. Integrable Models And The Toda Lattice Hierarchy by B.M.Sodermark
17. Perspectives Of Light-Front Quantized Field Theory -Some New Results by Prem P Srivastava
18. Gauge Symmetry In Chiral Electrodynamics by D.S.Kulshreshtha
19. Towards A Unified Description Of The Four Interactions In Terms Of Dirac-Bergmann  
Observables by L.Lusanna



# 14. Euclidean Methods In Quantum Field Theory

R. Ramanathan

Department of Physics and Astrophysics  
Univeristy of Delhi, Delhi -110007 (India)

## Abstract

In this article, some results from the rich and interesting formulation of Quantum field theory in Euclidean space-time are presented with a view to bring out the basic ideas of the Euclidean formulation. In particular, its rather surprising relationship with the "Stochastic Mechanics" of Nelson and the fundamental "Feynman-Kac" formulae which relate the Green's functions with the corresponding functional integrals, will be emphasized. The last section of this article briefly reviews a powerful Computational programme of Parisi and Wu which uses the Euclidean and Stochastic processes as a purely auxiliary tool in Quantum field theory. s.

## 1 Introduction

Quantum field theory in Minkowski space is beset with problems of positivity and finiteness of the norms involved in various Computations of physical quantities, which require ingenious artifices to overcome these pathological problems. One of the many "recipes" for avoiding such problems is the replacement of the usual description of quantum fields in Minkowski space-time with a description in some auxiliary Euclidean space. The resulting theory is often called the Euclidean quantum field theory in the literature.

In the early fifties Wick [1] used an Euclidean method while dealing with the Bethe-Salpeter equation. Schwinger [2] was one of the first to point out the feasibility of Euclidean formulation in a more comprehensive way as a "possible arena for the future development of quantum field theory" [3]. The correspondence between the Q.F.T formulated in Minkowski space and the Euclidean space can be achieved by making an analytical continuation of the points in Minkowski space of a vacuum expectation value of product of quantum fields, to the point of complex-space-time where the space coordinates are real and the time coordinate is purely imaginary (Schwinger points). The possibility of this analytical continuation is based on the work of Weightman and followers [4], on the general structure of quantum field theory, dictated by physical principles. The basic Covariance group of the Euclidean theory is the inhomogeneous Euclidean group (rotations and translations) unlike the poincare group (rotation, boost and translation) of the Minkowski theory. The structural simplification in going from the Minkowski to the Euclidean framework comes from the replacement of the indefinite metric of the Lorentz group (with the related hyperbolic problem). The resulting Euclidean theory is described by the so-called Schwinger functions, which are essentially wightman functions evaluated at Schwinger points. Symanzik [5] realised that Euclidean fields have a well defined dynamical structure, and showed that for the case of Bosons they are naturally defined as commutative random variables, while their dynamical structure shares a strong analogy with classical statistical mechanics. The mathematical foundation of Euclidean field theory was firmly laid by Nelson [6], who isolated the crucial Markov property (a Markov process is a stochastic process whose transition probability densities are independent of the past and future histories of the process, and solely depend on the present) as peculiar to the Euclidean field. He further established a Euclidean covariant Feynman-Kac path integral formula and showed the way to reconstruct a Wightman theory corresponding to a given Euclidean theory.

In order to appreciate the basic ideas of Euclidean field theory, let us consider a free scalar quantum field  $(\phi(x))$ , where  $x=\{x^\mu\}=(x^0 \equiv t; \mathbf{x})$  of mass ' $m$ ', completely defined through its vacuum expectation values (Wightman functions or Covariant Green's functions) [4].



$$W(x_1 \dots x_{2n}) = \langle 0 | \phi(x_1) \dots \phi(x_{2n}) | 0 \rangle = \sum_{j=2}^{2n} W(x_1, x_j) W(x_2 \dots x_j \dots) \quad (1)$$

with the expectation of an odd number of fields vanishing. The two-point function is given by :

$$W(x, y) = W(x - y) = (2\pi)^{-3} \exp[-iw(\mathbf{k})(x_0 - y_0)] \otimes \exp[i\mathbf{k} \cdot (\mathbf{x} - \mathbf{y})] d^3\mathbf{k} / 2w(\mathbf{k}) \quad (2)$$

where

$$w(\mathbf{k}) = \sqrt{(\mathbf{k}^2 + m^2)} \geq m \quad (3)$$

If we perform the analytic continuation (wick rotation) on  $x_0, y_0$  to Schwinger points.

$$x_0 = -ix_4; \quad y_0 = -iy_4; \quad x_4 \geq y_4 \quad (4)$$

Then we get the Schwinger function

$$S(x, y) = S(x - y) = (2\pi)^{-3} \int \exp\{-w(\mathbf{k})(x_4 - y_4)\} \otimes \exp\{i\mathbf{k} \cdot (\mathbf{x} - \mathbf{y})\} d^3\mathbf{k} / 2w(\mathbf{k}) \quad (5)$$

where  $(x, y)$  are the Euclidean coordinates  $(\mathbf{x}, x_4)$  and  $(\mathbf{y}, y_4)$ .

Using the identity

$$(1/2\pi) \exp\{ik_4(x_4 - y_4)\} (k_4^2 + w^2(\mathbf{k}))^{-1} dk_4 = 1/2w(\mathbf{k}) \exp\{-w(\mathbf{k}) | x_4 - y_4 | \} \quad (6)$$

we can rewrite (5) in the form

$$S(x, y) = (2\pi)^{-4} \exp\{ik \cdot (x - y)\} (k^2 + m^2)^{-1} d^4k \quad (7)$$

where

$$x, y \in R^4; \quad k \cdot x = k_4 x_4 + \mathbf{k} \cdot \mathbf{x}; \quad k^2 = k_4^2 + \mathbf{k}^2; \quad d^4k = dk_4 d^3\mathbf{k}$$

Clearly (7) has a validity for all  $x_4, y_4$  and has explicit Euclidean covariance in the 4-Euclidean space. Exploiting the positive definiteness of (7), Euclidean fields are introduced as Gaussian random fields with mean zero and covariance

$$E(\phi(x)\phi(y)) = S(x, y) \quad (8)$$

Therefore we see that Euclidean theory can be described through stochastic (Commutative) fields as opposed to the non-commutative Heisenberg fields of the Minkowski theory.

The above illustrative example for the free scalar field can be extended to both free and interacting vector and spinor fields as discussed in detail in [3].

## 2 Nelson's Stochastic Mechanics

There is a deep and surprising connection between the Euclidean field theory formulated in imaginary time and the stochastic mechanics of Nelson [7], which arose from an interesting derivation of Schrodinger equation from a classical, but stochastic dynamical law resembling Newtonian mechanics. Whatever view one may hold about the foundational aspects of this derivation, insofar as it offers an 'objective-realist' basis for quantum mechanics, it is interesting from both mathematical and applicational viewpoints. In brief Nelson's derivation may be summarized as follows. The probability density

$$\rho(x, t) = |\psi(x, t)|^2 \quad (9)$$

where  $\psi$  is the Schroedinger wave function in  $R^4$  satisfying

$$i\hbar\partial_t\psi = -\frac{\hbar^2}{2m}\nabla^2\psi - V\psi \quad (10)$$

obeys the continuity equation :

$$\dot{\rho} = -\nabla \cdot \vec{j} \quad (11)$$

where the current  $\vec{j}$  is

$$\vec{j} = -i(\hbar/2m)\bar{\psi}\nabla\psi - \hbar/m\rho\nabla(\text{Im} \ln \psi) \quad (12)$$

It is possible to associate a Markov process to  $\rho$  described by a transition probability density  $p(\mathbf{x}, t; \mathbf{x}', t')$  such that

$$\rho(\mathbf{x}, t)_{t,t'} = \int P(\mathbf{x}, t; \mathbf{x}', t')\rho(\mathbf{x}', t')d\mathbf{x}' \quad (13)$$

will be obey a Kolmogorov-Fokker Planck equation

$$\dot{\rho} = (v/2)\nabla^2\rho - \nabla \cdot (\vec{b}\rho) \quad (14)$$

Eqs (12) and (14) are compatible if

$$v = \hbar/m \quad (15)$$

and

$$\nabla^2 |\psi|^2 = -i\nabla\bar{\psi} \cdot \nabla\psi + 2m/\hbar \nabla \cdot (\vec{b} |\psi|^2) \quad (16)$$

One further condition

$$\nabla \times \vec{b} = 0 \quad (17)$$

leads to the solution

$$\vec{b} = (\hbar/m)\nabla[\ln |\psi| + \arg \psi] \quad (18)$$

We thus have two velocities, the current velocity  $\vec{u}$  defined as

$$\vec{u} = (\hbar/m)\nabla \ln |\psi| \quad (19)$$

and the osmotic velocity

$$\vec{v} = (\hbar/m)\nabla \arg \psi \quad (20)$$

The Schroedinger equation leads to the following equations of motion for  $\vec{u}$  and  $\vec{v}$  :

$$\dot{\vec{u}} = -(\hbar/2m)\nabla(\nabla \cdot \vec{v}) - \nabla(\vec{u} \cdot \vec{v}) \quad (21)$$

$$\dot{\vec{v}} = -(1/m)\nabla(\vec{u} \cdot \vec{v}) + 1/2\nabla(u^2 - v^2) + (\hbar/2m)\nabla^2\vec{u} \quad (22)$$

An equivalent description of the stochastic process is given by the Langevin equation [8]

$$dx = bdt + \sqrt{(\hbar/2m)}dw \quad (23)$$

where  $dw$  is the increment of the wiener process described by

$$\langle dw_i \rangle = 0; \quad \langle dw_i dw_j \rangle = 2\delta_{ij}\Delta t \quad (24)$$

and the condition that  $\langle dw_i \rangle$  are Gaussian mean values which are independent for different times. For the ground state of the Hamiltonian

$$H = -(\hbar^2/2m\nabla^2 + V) \quad (25)$$

it has no nodes, and we see immediately that  $\vec{v} = 0$ ; the process is stationary (this in fact is true for any eigenstate of the Hamiltonian [8]).

As a simple example let us consider the ground state of the harmonic oscillator [9]

$$\psi_0 = (2\pi\sigma)^{-1/4} \exp(-x^2/4\sigma); \quad \sigma = \hbar/(2mw) \quad (26)$$

which leads to the "drift"

$$b = -(\hbar/2m\sigma)x = -wx \quad (27)$$

The Kolmogorov-Fokker-Planck (K.F.P.) equation becomes

$$\rho = (\hbar/2m)d^2/dx^2 \rho + w\rho + wx d\rho/dx \quad (28)$$

which has a solution

$$\rho(x, t) = \int P(x^1, 0) p(x, x^1; t) dx^1$$

with the transition probability

$$P(x, x^1; t) = (2\pi\bar{\sigma}(t))^{-1/2} \exp[1/2\bar{\sigma}(t)(x - x^1 e^{-wt})^2] \quad (29)$$

where

$$\bar{\sigma}(t) = (1 - e^{-2wt}) \quad (30)$$

One thus observes a most surprising result

$$\langle x(0)x(t) \rangle = \sigma e^{-w|t|} \quad (31)$$

which looks very "Euclidean". More generally we observe that the ground state process leads to the K.F.P. equation

$$\rho = (\hbar/2m)[\nabla^2 \rho - (\nabla \ln \rho_0)\rho] \quad (32)$$

which has the stationary solution  $\rho = \rho_0 \ln |\psi_0|^2$ . It turns out that quite generally

$$\langle x(0)x(h) \rangle = (\psi_0, x e^{-H|h|} x \psi_0) \quad (33)$$

(33) is a remarkable formula because it links a "real time" object on the left-hand-side to an "imaginary time" quantity on the right-hand-side.

This general connection was noted by Guerra and Ruggiero [9]: "Euclidean Quantum Mechanics (or Field theory) is the ground state process of Stochastic Mechanics". The relation (33) is true only because the "real-time" left-hand-side is not accessible to measurement. We also note that the K.F.P. equation has the formal solution

$$\rho(t) = e^{-ht} \rho(0) = \psi_0 e^{-Ht} \psi_0^{-1} \rho(0) \quad (34)$$

where  $H = -(\hbar^2/2m)\nabla^2 + V$  and  $b = 2\nabla \ln |\psi_0|$

The above relation suggests a possible application of stochastic mechanics. We have to first obtain a trial ground state  $\psi_0$  and thence  $b$ . We then use this to obtain the ground state energy levels using Langevin or Monte Carlo methods [10].

### 3. Euclidean Path integrals and Feynman-Kac Formulae

As we saw earlier, the Euclidean approach uses the positivity of the Hamiltonian to achieve the 'Euclidean' rotation from the time evolution  $\exp iHt$  of 'real time' quantum mechanics to  $\exp -Ht$  of the 'imaginary time' Euclidean field theory. In this procedure the Stochastic processes play a purely auxiliary role since they do not take place in "real time". When we have the Hamiltonian

$$H = -\nabla^2 + V \quad (35)$$

where we set  $\hbar = 2m = 1$ , The Feynman-Kac formula gives a representation for the integral Kernel (with respect to the Lesbesgue measure  $dx$ ) as follows:

$$e^{-tH(x,y)} = \int dP_{xy}^t(w) \exp \left[ - \int_0^t V(w(\tau)) d\tau \right] \quad (36)$$

Here  $dP_{xy}^t(w)$  is the conditional Wiener measure for Brownian paths starting at  $x$  and ending in  $y$  after time  $t$ . The most straightforward way to derive (36) is based on the famous "Trotter product formula"

$$e^{-tH} = \lim_{n \rightarrow \infty} (e^{t\Delta/n} e^{-tv/n})^n \quad (37)$$

where the limit holds in the strong sense. The form of the Feynman-Kac formula found most frequently gives instead the solution of the imaginary time Schroedinger equation

$$.\dot{\psi} = (\nabla^2 - V)\psi = H\psi \quad (38)$$

by

$$\psi(x, t) = \int dP^t(w) \exp \left[ - \int_0^t V(x + w(\tau)) d\tau \right] \psi(x + w(t), 0) \quad (39)$$

where  $dP^t(w)$  is now the standard Wiener measure for paths starting at the origin.

It is now easy to extract the Euclidean path integral from the Feynman-Kac formula for the Schwinger functions, for instance

$$\begin{aligned} \langle x(t_1) \dots x(t_n) \rangle &= (\psi_0, x e^{(t_2-t_1)H} \dots x e^{(t_n-t_{n-1})H} \psi_0) \\ &= \lim_{T \rightarrow \infty} (1/Z_T) \int dx + \int dP_{xx}^t(w) \\ &\quad \times \exp \left[ - \int_0^t V(w(\tau)) d\tau \right] w(t_1) \dots w(t_n) \end{aligned} \quad (40)$$

where  $T \geq t_1 \geq t_2 \dots \geq t_n \geq 0$  and

$$Z_T = \int dx \int dp^{T*} \exp \left[ - \int_0^T V(w(\tau)) d\tau \right] \equiv \text{Tr}[e^{-TH}] \quad (41)$$

Equation (40) can be recast formally as

$$\langle x(t_1) \dots x(t_n) \rangle = \lim_{T \rightarrow \infty} (1/Z_T) \int e^{-S_T x(t_1) \dots x(t_n)} \prod_{t=0}^T dx(t) \quad (42)$$

where  $S_T$  is the Euclidean action in time  $T$  and

$$Z_T = \int e^{-S_T} \prod_{t=0}^T dx(t)$$

This is so because the conditional Wiener measure corresponds to the expression.

$$.1/N_T \exp \left[ - \int_0^T w((\tau))^2 d\tau \right] \prod_{t=0}^T dw(t) \quad (43)$$

It is well known Gaussian integration techniques that facilitate easy evaluation of the integrals involved in the formula for Schwinger functions, that the Euclidean path integrals are far easier to handle for any given Euclidean action. In the foregoing, rather brief introduction to Euclidean field theory, we have confined ourselves to only the formal aspects. All the results of conventional Minkowski fields can be reproduced much more elegantly in Euclidean field theory, for more detail on applicational aspects one can refer to Simon [11].

## 4. Stochastic Field Theory and Euclidean Field Theory

It is natural to seek extension of Nelson's stochastic mechanics to systems with infinite degrees of freedom, which are fields. We only indicate the directions along which such an extension can be achieved without going into the applicational aspects which would entail a voluminous extension of the text of this Chapter. As such an elaborate treatment of the Field theoretic extension to Nelson's stochastic mechanics is not on our agenda; we shall only give a brief and introductory survey of this area. We shall lay emphasis on two approaches namely Stochastic Field theory and Euclidean Field theory which are surprisingly interrelated. As we have seen earlier, one great advantage of these approaches is the relative ease in computing the correlation functions because of the facility offered by real gaussian integrals as against the complex integrals with operator valued functions occurring in the conventional quantum field theory. This is the main positive aspect of these approaches to field theory.

### 4.1 Stochastic Mechanics of Free Scalar Field

The first step in the extension of the Nelsonian approach to quantum mechanics to encompass fields is the stochasticisation of the free scalar field. Consider the free scalar field Hamiltonian in natural units ( $\hbar = c = 1$ )

$$H = 1/2 \int [\partial_t \phi^2 + \nabla \phi^2 + m_0^2 \phi^2] d^3 x; \quad x \in R^3 \quad (44)$$

And the equation of motion

$$\partial_t^2 \phi - \nabla^2 \phi + m_0^2 \phi = 0 \quad (45)$$

Let  $\{u_n(x)\}$ , ( $n$  positive integers), be a complete set of basis vectors with box normalisation in a finite box of volume  $V$ . We then have the normalisation

$$\int_V u_n(\vec{x}) u_{n'}(\vec{x}) d^3 x = \delta_{nn'} \quad (46)$$

and the completeness relation

$$\sum_n u_n(\vec{x}) u_n(\vec{x}') = \delta_V^3(\vec{x} - \vec{x}') \quad (47)$$

and the harmonic oscillator equation

$$\nabla^2 u_n(\vec{x}) = -k_n^2 u_n(\vec{x}) \quad (48)$$

Going over to the infinite volume limit, we have

$$\delta_V(\vec{x} - \vec{x}') \rightarrow \delta(\vec{x} - \vec{x}') = (2\pi)^{-3} \int \exp[i\vec{k} \cdot (\vec{x} - \vec{x}')] d^3 \vec{k} \quad (49)$$

Expanding  $\phi(\vec{x}, t)$  in the form

$$\phi(\vec{x}, t) = \sum_{n=0}^{\infty} u_n(\vec{x}) q_n(t) \quad (50)$$

where  $q_n(t)$  are independent variables, each satisfying a harmonic oscillator equation (The free field is an infinite set of harmonic oscillators). Then

$$\partial_t^2 q_n(t) + \omega_n^2 q_n(t) = 0; \quad \omega_n^2 = m_n^2 + k_n^2 \quad (51)$$

where the Hamiltonian is given by

$$H_n = 1/2 p_n^2 + 1/2 \omega_n^2 q_n^2 \quad (52)$$

By promoting  $q_n(t)$  to independent Gaussian Markov processes with averages given by

$$\langle q_n(t) \rangle = 0; \quad \langle q_n(t), q'_n(t') \rangle = \delta_{nn'} (2w_n)^{-1} \exp[-w_n |t - t'|] \quad (53)$$

Consequently from eqn.(50),  $\phi(\vec{x}, t)$  is also raised to a Gaussian Markov process with averages

$$\langle \phi(x, t) \rangle = 0; \quad \langle \phi(\vec{x}, t) \phi(\vec{x}', t') \rangle = \sum_{n=0} u_n(\vec{x}) u_n(\vec{x}') (2w_n)^{-1} \exp[-w_n |t - t'|] \quad (54)$$

in a finite box  $V$ . In the limit of  $V \rightarrow R^3$ , utilising (47) and (49) we can write the two-point correlation of  $\phi$  in the form

$$\langle \phi(\vec{x}, t) \phi(\vec{x}', t') \rangle = \int \exp[-\sqrt{k^2 + m_0^2} |t - t'|] \frac{d^3 \vec{k} \exp\{i \vec{k} \cdot (\vec{x} - \vec{x}')\}}{2(2\pi)^3 \sqrt{k^2 + m_0^2}} \quad (55)$$

Now consider the stochastic differential equations for  $q_n(t)$ , viz.,

$$dq_n(t) = -w_n q_n(t) dt + dw_n \quad (56)$$

The noise  $dw_n$  is normalised as

$$dw_n dw_{n'} = \delta_{nn'} \delta_t \quad (57)$$

If we use (56) in (50) and go to the infinite volume limit we get the stochastic differential equation for the field  $\phi_n(\vec{x}, t)$

$$d\phi(\vec{x}, t) = -\sqrt{-\nabla^2 + m_0^2} \phi(\vec{x}, t) dt + dw(\vec{x}, t) \quad (58)$$

Thus the Nelsonian framework can be extended to the free scalar field quantisation (i.e. association of a Markovian stochastic process to the quantum state of a dynamical system). In particular the ground state process for a free scalar field is found to be a Gaussian Markov field with the two-point correlation given by eqn.(55).

## 4.2 Stochastic Mechanics of the e-m Field

The Maxwell field in free space is a mechanical system with dynamical variables given by the electric field  $\vec{E}(x, t)$  and the magnetic field  $\vec{B}(x, t)$ . The Hamiltonian is

$$H = 1/2 \int [[E^2(x, t) + B^2(x, t)] d^3 \vec{x} \quad (59)$$

where the equation of motion are

$$\partial_t \vec{B} = -\nabla \times \vec{E}; \quad \partial_t \vec{E} = \nabla \times \vec{B} \quad (60)$$

and

$$\nabla \cdot \vec{B} = \nabla \cdot \vec{E} = 0 \quad (61)$$

Through the analogical extension of the standard methods of Nelson's stochastic mechanics, we may set up the stochastic differential equations for the Maxwell field by promoting  $\vec{B}(\vec{x}, t)$  to a Gaussian Markov field and  $\vec{E}(\vec{x}, t)$ , as the drift fields of stochastic mechanics as  $\vec{E}_{(\pm)}(\vec{B}, \vec{x}, t)$ , such that the following stochastic differential equations are satisfied:

$$\begin{aligned} d\vec{B}(\vec{x}, t) &= -\nabla \times \vec{E}_{(\pm)}(\vec{B}, \vec{x}, t) dt + dw(\vec{x}, t) \\ (D_{(\pm)} \vec{B})(\vec{x}, t) &= -\nabla \times \vec{E}_{(\pm)}(\vec{B}, \vec{x}, t) \\ \nabla \times \vec{B} &= 1/2 [D_{(+)} \vec{E}_{(-)} + D_{(-)} \vec{E}_{(+)}] \end{aligned} \quad (62)$$

The noise, however, can no longer be pure white noise as in the case of single particle dynamics (see earlier chapters), in order to take care of the transversality condition (61)

$$dW_\alpha(\vec{x}, t) dW_\beta(\vec{x}, t) = (2\pi)^{-3} \int \exp[i\vec{k} \cdot (\vec{x} - \vec{x}')] \times [k^2 \delta_{\alpha\beta} - k_\alpha k_\beta] d^3 \vec{k} \quad (63)$$

The above set of stochastic differential equation may be used to compute the two-point correlation function for the Maxwell field. It has been shown that it indeed yields results which are identical with the conventional quantum electro dynamics. As already mentioned, we will not delve into the intricacies of this aspect of stochastic field theory as it is beyond the scope of this article and we will only urge the interested reader to the copious literature on the subject [9, 11].

## 5. The Parisi-Wu Stochastic Quantisation

The stochastic quantisation scheme of Parisi and Wu [12] exhibits a Euclidean Quantum field  $\psi(x)$  as the stationary limit with respect to a fictitious time  $\tau$  (like the computer time of a Monte-Carlo simulation) of the stochastic relaxation process defined for  $\tau \geq 0$  by a generalised Langevin equation

$$\partial_\tau \psi(\tau x) = -\delta S[\psi(\tau x)] / \delta \psi(\tau, x) + \eta(\tau, x) \quad (64)$$

Here  $x = \{x^\nu\}$ ,  $\nu = 0, \dots, 3$ ;  $S$  is the classical Euclidean action and  $\eta$  is a Gaussian white noise with correlation function.

$$\langle \eta(\tau, x) \eta(\tau', x') \rangle_\eta = 2\pi \delta(\tau - \tau') \delta^4(x - x') \quad (65)$$

The Euclidean Green's functions (Schwinger functions) are obtained as the "equilibrium limit" of the correlation functions of the process :

$$\langle \psi(x_1) \dots \psi(x_n) \rangle = \lim_{\tau_n \rightarrow \infty} \langle \psi(\tau_1, x_1) \dots \psi(\tau_n, x_n) \rangle_n \quad (66)$$

A most interesting aspect of the Parisi-Wu method for continuum field theories is that in gauge theories the perturbative calculation of the left hand side of eqn. (66) may be based on the classical action  $S$  alone, i.e. no gauge-fixing term and associated Faddey-Popov ghosts are necessary.

This approach thus uses stochasticity as an auxiliary tool, but here the stochastic processes take place in the fictitious 'fifth dimension' (. To make clear the link between the Feynman-Kac formula of the earlier section, we shall see that it leads to Euclidean Functional integrals. The idea is now to set up a stochastic process possessing the Euclidean functional measures as its unique equilibrium measure. Clearly this does not specify the process uniquely. But the simplest choice is based on a Langevin equation like (64).

We now only have to make the identification

$$|\psi_0|^2 = Z^{-1} \exp[-S/\hbar] \quad (67)$$

where  $Z$  is a normalizing factor

$$\vec{b} = \nabla S / \hbar \quad (68)$$

where ' $\nabla$ ' now really means a functional gradient. At this formal level there is no difference between ordinary quantum mechanics and quantum field theory, so we assume that  $S$  is a functional of some Euclidean fields symbolised as  $\psi$ .

The Kolmogorov-Fokker-Planck equation now reads (with a suitably rescaled time) as

$$\partial_t \rho = \hbar^2 \nabla^2 \rho + \nabla \cdot (\nabla S) \rho \quad (69)$$

. and the Langevin equation

$$d\phi = \nabla S dt + \hbar dw \quad (70)$$

$dw$  is now a higher dimensional Wiener process, one for each space-time point and field component. For a scalar field  $\phi$  we may take

$$\langle dw(x)dw(y) \rangle = 2\delta(x-y)dt \quad (71)$$

We can also determine a potential  $V$  belonging to the ground state wave function

$$\psi_0 = Z^{1/2} \exp[-1/2S/\hbar]$$

It is

$$V = \hbar^2[\nabla^2\psi_0]/\psi_0 = -\hbar\nabla^2 S + (\nabla S)^2 \quad (72)$$

where the ' $\nabla$ ' operators are to be interpreted as functional derivatives. So one can write a Feynman-Kac formula leading to a functional integral with one extra dimension. Formally its density is given by

$$\exp[-\tilde{S}/\hbar]; \quad \tilde{S} = 1/4 \int \phi^2 d\tau + \int V d\tau \quad (73)$$

The "super euclidean" models arising in this way here exhibit interesting supersymmetry as shown by Parisi, Sourlas and Gozzi [13].

Stochastic quantisation of Parisi-Wu can be used to construct a perturbation expansion [12]. It is also possible to use the auxiliary time for regularization by replacing the Wiener process by a suitable non-Markovian process; this amounts to replacing  $\int \phi^2 d\tau$  in (73) by  $\{\phi, C^{-1}\phi\}$  where  $C^{-1}$  is a suitable operator. In view of applications to gauge theory it is useful to choose  $C^{-1}$  "local in space-time".

Our intention in this section was mainly to highlight some essential features of the Parisi-Wu scheme which is a powerful computational scheme using stochasticity as an auxiliary tool unlike the Euclidean and stochastic approaches to field theory which use stochasticity as a core ingredient of the formulation. We also note in passing that the field theories dealt here are purely at zero temperature and should be distinguished from finite temperature field theories or the thermo-field theories, although in those formulations too, Euclidean structures for the action are postulated; see Article by Ashoke Das in this Book.

## References

- [1] G.C.Wick, Phys. Rev-96 (1954) 1124.
- [2] J.Schwinger, in Proc. Of the 1958 Conf. on High Energy Physics of CERN. Ed. B. Ferretti (CERN. Gexeva. 1958).
- [3] B.Simon, the P(2) Euclidean (Quantum) Field Theory (Princeton N.J. 1974).
- [4] R.Streater and A.S.Wightman, P.C.T, Spin and Statistics, and all that (Benjamin, N.J 1964).
- [5] K.Symanzik, local Quantum Theory .ed. R.Jost (Academic press, New York, 1969)
- [6] E.Nelson, Journ. Funct. Analysis 12 (1973) 97; 12 (1973) 211.
- [7] E.Nelson, "Dynamical theories of Brownian Motion (Princeton University press, Princeton, 1967).
- [8] E.Nelson, Phys. Rev. 150 (1966) 1079, Also see [7]; R.Ramanathan, Phys. Scripta 34, 365 (1986).
- [9] F.Guerra, Phys. Rep. 77 (1981) 263.
- [10] G.Jana-Lasinio, "Stochastic Process and Quantum Mechanics", Ecole Poly-technique report, June 1983.



- [11] B.Simon, Functional integration in Quantum physics (Academic Press, N.Y. - San Francisco - London, 1979).
- [12] G.Parisi and Y.S.Wu, Sci. Sinica, 24 (1981) 483.
- [13] E.Gozzi, Phys. Lett. 130B (1983) 83; Phys. Rev., D28 (1983); G.Parisi and N.Sourlas, Nucl. Phys., B206 (1982) 321.

# 15. Topics in Finite Temperature Field Theory

Ashok Das \*

Department of Physics and Astronomy,  
University of Rochester,  
Rochester, New York, 14627

## Abstract

We discuss a few selected topics in finite temperature field theory.

## 1 Introduction

Studies of physical systems at finite temperature have led, in the past, to many interesting properties such as phase transitions, blackbody radiation etc. However, the study of complicated quantum mechanical systems at finite temperature has had a systematic development only in the past few decades. There are now well developed and well understood formalisms to describe finite temperature field theories, as they are called. In fact, as we know now, there are three distinct, but equivalent formalisms [1-3] to describe such theories and each has its advantages and disadvantages. But, the important point to note is that we now have a systematic method of calculating thermal averages perturbatively in any quantum field theory.

This, of course, has led to a renewed interest in the study of finite temperature field theories for a variety of reasons. We can now study questions such as phase transitions involving symmetry restoration in theories with spontaneously broken symmetry [4]. We can study the evolution of the universe at early times which clearly is a system at high temperature. More recently, even questions such as the chiral symmetry breaking phase transition or the confinement-deconfinement phase transition in QCD [5-6] have drawn a lot of attention in view of the planned experiments involving heavy ion collisions. This would help us understand properties of the quark-gluon plasma better.

The goal of this article is to share, with the readers, some of the developments in finite temperature field theories in the recent past and the plan of the article is as follows. In the next section, we will describe some basic ideas behind describing a quantum mechanical theory in terms of path integrals [7]. This is the approach which generalizes readily to the study of finite temperature field theory. In section 3, we will discuss one of the formalisms, in fact, the oldest one, of describing finite temperature field theory. This goes under the name of the imaginary time formalism or the Matsubara formalism [1, 5, 8-10]. In this description, the dynamical time is traded in for the temperature. In contrast, the real time formalisms of finite temperature field theory contain both time and temperature. In section 4, we discuss one of the real time formalisms known as thermo field dynamics [3, 10-11]. This is an ideal description to understand operator related issues involving finite temperature field theories although it has a path integral representation which is quite nice for calculations as well. The other real time formalism, which is much older and is known as the closed time path formalism [2, 10, 12], is described in section 5. This formalism is very nice because it describes both equilibrium and non-equilibrium phenomena, at finite temperature, with equal ease. Temperature leads to many subtle features in field theories. In section 6, we discuss one such subtlety, namely, how one needs a generalization of the Feynman combination formula to perform calculations at finite temperature [13]. In section 7, the issue of large gauge invariance

---

\*Email: das@hep.pas.rochester.edu

is discussed within the context of a simple quantum mechanical model [14-15]. In section 8, we discuss in some detail how temperature can lead to breaking of some symmetries like supersymmetry [16] (Temperature normally has the effect of restoring symmetries). Finally, we present a brief conclusion in section 9. The subject of finite temperature field theories is quite technical and to keep the contents simple, we have chosen, wherever possible, simple, quantum mechanical models to bring out the relevant ideas. Finally, we would like to note that there are many works in the literature and the references, at the end, are only representative and are not meant to be exhaustive in any way.

## 2 Path Integrals at Zero Temperature

In studying a quantum mechanical system or a system described by a quantum field theory, we are basically interested in determining the time evolution operator. In the standard framework of quantum mechanics, one solves the Schrödinger equation to determine the energy eigenvalues and eigenstates simply because the time evolution operator is related to the Hamiltonian. There is an alternate method for evaluating the matrix elements of the time evolution operator which is useful in studying extremely complicated physical systems. This goes under the name of path integral formalism [7, 17-18].

In stead of trying to develop the ideas of the path integral formalism here, let us simply note that, for a bosonic system described by a time independent quantum mechanical Hamiltonian, the transition amplitude can be represented as (The subscript  $H$  denotes the Heisenberg picture.)

$${}_H\langle x_f, t_f | x_i, t_i \rangle_H = \langle x_f | e^{-\frac{i}{\hbar} H(t_f - t_i)} | x_i \rangle = \int \mathcal{D}x e^{\frac{i}{\hbar} S[x]} \quad (1)$$

There are several comments in order. First, the transition amplitude is nothing other than the matrix element of the time evolution operator in the coordinate basis. Second, the integral on the right hand side is known as a path integral. It is an integral over all possible paths connecting the initial coordinate  $x_i$  and the final coordinate  $x_f$  which are held fixed. The simplest way to evaluate such an integral is to divide the time interval of the path between  $x_i$  and  $x_f$  into  $N$  intervals of equal length. Integrating over all possible values of the coordinates of the intermediate points (which are ordinary integrals) and taking  $N \rightarrow \infty$  such that the time interval is held fixed is equivalent to integrating over all possible paths. Finally, the action  $S[x]$  in the exponent of the integrand is nothing other than the classical action for the bosonic system under study. This is true for most conventional physical systems where the Hamiltonian depends quadratically on the momentum. If this is not the case (and there are some cases where it is not), the right hand side of (1) needs to be modified. However, for most systems that we will discuss, we do not have to worry about this fine point.

The advantage of the path integral is that while the left hand side involves quantum mechanical operators, the right hand side is described only in terms of classical variables and, therefore, the manipulations become quite trivial. Furthermore, the transition amplitude defined in eq. (1) can be generalized easily to incorporate sources and this allows us to derive various Greens functions of the theory in a very simple and straightforward manner. As an example, let us simply note here that for a harmonic oscillator, the action is quadratic in the dynamical variables, namely,

$$S[x] = \int_{t_i}^{t_f} dt \left[ \frac{1}{2} m \dot{x}^2 - \frac{1}{2} m \omega^2 x^2 \right]$$

and, in this case, the path integral can be exactly evaluated and has the form [7]

$$\begin{aligned} \langle x_f | e^{-\frac{i}{\hbar} H T} | x_i \rangle &= \int \mathcal{D}x e^{\frac{i}{\hbar} S[x]} \\ &= \left( \frac{m\omega}{2\pi i \hbar \sin \omega T} \right)^{\frac{1}{2}} e^{\frac{i}{\hbar} S[x_{cl}]} \end{aligned} \quad (2)$$

Here, we have defined  $T = t_f - t_i$ .  $S[x_{cl}]$  represents the action associated with the classical trajectory (satisfying the Euler-Lagrange equation) and has the form

$$S[x_{cl}] = \frac{m\omega}{2\sin\omega T} [(x_i^2 + x_f^2) \cos\omega T - 2x_i x_f] \quad (3)$$

The path integrals can also be extended to quantum mechanical systems describing fermionic particles. However, one immediately recognizes that there are no classical variables which are fermionic. Therefore, in order to have a path integral description of such systems in terms of classical variables, we must supplement our usual notions of classical variables with anti-commuting Grassmann variables [19]. With this, for example, we can write a classical action for the fermionic oscillator as

$$S[\psi, \bar{\psi}] = \int_{t_i}^{t_f} dt (i\bar{\psi}\dot{\psi} - \omega\bar{\psi}\psi) \quad (4)$$

Here  $\psi$  and  $\bar{\psi}$  are anti-commuting Grassmann variables and in the quantum theory, as operators, can be identified with the fermionic annihilation and creation operators respectively. The action in eq. (4) is also quadratic in the variables much like the bosonic oscillator and the path integral for the fermionic oscillator can also be exactly evaluated giving [7]

$$\begin{aligned} \langle \psi_f, \bar{\psi}_f | e^{-\frac{i}{\hbar}HT} | \psi_i, \bar{\psi}_i \rangle &= \int \mathcal{D}\bar{\psi} \mathcal{D}\psi e^{\frac{i}{\hbar}S[\psi, \bar{\psi}]} \\ &= e^{\frac{i\omega T}{2}} e^{(e^{-i\omega T} \bar{\psi}_f \psi_i - \bar{\psi}_f \psi_f)} \end{aligned} \quad (5)$$

In a quantum field theory, we are often interested in evaluating time ordered correlation functions in the vacuum because the S-matrix elements can be obtained from such Greens functions. These can be derived in a natural manner from what is known as the vacuum to vacuum transition functional which can be obtained from the transition amplitude in eq. (1) in a simple manner and also has a path integral representation of the form

$$\lim_{T \rightarrow \infty} \langle 0 | e^{-\frac{i}{\hbar}HT} | 0 \rangle = \int \mathcal{D}x e^{\frac{i}{\hbar}S[x]} \quad (6)$$

where

$$S[x] = \int_{-\infty}^{\infty} dt L(x, \dot{x}) \quad (7)$$

Furthermore, the path integral in eq. (6) has no end-point restriction unlike in eq. (1). This vacuum to vacuum transition amplitude is also commonly denoted by  $\langle 0|0 \rangle$  with the limiting process understood. We note here that an analogous formula also holds for fermionic systems.

The vacuum to vacuum amplitude in the presence of a source has the form

$$Z[J] = \langle 0|0 \rangle_J = \int \mathcal{D}x e^{\frac{i}{\hbar}S[x, J]} \quad (8)$$

where

$$S[x, J] = S[x] + \int_{-\infty}^{\infty} dt J(t)x(t) \quad (9)$$

Here  $J(t)$  is a classical source and it can be easily checked that, in the limit of vanishing source, the functional derivatives of  $Z[J]$  give rise to time ordered Greens functions in the vacuum.

With this very brief review of the path integral description for zero temperature quantum mechanical theories, we are now ready to describe the different formalisms available to study quantum mechanical systems at finite temperature.

### 3 Imaginary Time Formalism

The properties of a quantum mechanical system, at finite temperature, can also be given a path integral description. There are various, but equivalent ways of doing this. Of the different formalisms available to study a quantum mechanical system at finite temperature, the imaginary time formalism is the oldest [1]. To appreciate this, let us recall some of the features of a statistical ensemble. A statistical ensemble in equilibrium at a finite temperature  $\frac{1}{\beta}$  (in units of Boltzmann constant) is described in terms of a partition function

$$Z(\beta) = \text{Tr } \rho(\beta) = \text{Tr } e^{-\beta \mathcal{H}} \quad (10)$$

Here  $\rho(\beta)$  is known as the density matrix (operator) and  $\mathcal{H}$  can be thought of as the generalized Hamiltonian of the system. If

$$\mathcal{H} = H$$

where  $H$  is the Hamiltonian of the system, we say that the ensemble is a canonical ensemble where the particle number is fixed and the system is allowed to exchange only energy with a heat bath. On the other hand, if

$$\mathcal{H} = H - \mu N$$

where  $N$  is the number operator, then, the ensemble is known as a grand canonical ensemble where the system can exchange not only energy with a heat bath, but can also exchange particles with a reservoir. The constant  $\mu$  is known as the chemical potential. In a statistical ensemble, of course, the important observables are the ensemble averages and, for any observable  $\mathcal{O}$ , they are defined as

$$\langle \mathcal{O} \rangle_\beta = \frac{1}{Z(\beta)} \text{Tr } \rho(\beta) \mathcal{O} \quad (11)$$

Let us also note here that since the partition function involves a trace, it leads to an interesting identity following from the cyclicity of the trace, namely, (we will assume from now on, unless otherwise specified, that  $\hbar = 1$ )

$$\begin{aligned} \langle \mathcal{O}_1(t) \mathcal{O}_2(t') \rangle_\beta &= \frac{1}{Z(\beta)} \text{Tr } e^{-\beta \mathcal{H}} \mathcal{O}_1(t) \mathcal{O}_2(t') \\ &= \frac{1}{Z(\beta)} \text{Tr } e^{-\beta \mathcal{H}} \mathcal{O}_2(t') e^{-\beta \mathcal{H}} \mathcal{O}_1(t) e^{\beta \mathcal{H}} \\ &= \frac{1}{Z(\beta)} \text{Tr } e^{-\beta \mathcal{H}} \mathcal{O}_2(t') \mathcal{O}_1(t + i\beta) \\ &= \langle \mathcal{O}_2(t') \mathcal{O}_1(t + i\beta) \rangle_\beta \end{aligned} \quad (12)$$

Such a relation is known as the KMS (Kubo-Martin-Schwinger) [20] relation which generalizes to all statistical ensemble averages and plays a crucial role in the study of finite temperature field theories.

It was observed quite early by Bloch [21] that the operator  $e^{-\beta \mathcal{H}}$  in the definition of the partition function is like the time evolution operator in the imaginary time axis. This is really at the heart of the imaginary time formalism. In fact, let us note that the canonical partition function can be written as (with the trace taken in the coordinate basis)

$$Z(\beta) = \int dx \langle x | e^{-\beta H} | x \rangle \quad (13)$$

It is clear now that if we identify  $T = -i\beta$  in eq. (1), then, we can give the partition function a path integral representation as ( $\hbar = 1$ )

$$Z(\beta) = \int \mathcal{D}x e^{-S_E[x]} \quad (14)$$

where  $S_E[x]$  is the Euclidean (imaginary time) action for the system defined over a finite time interval as

$$S_E[x] = \int_0^\beta dt L_E(x, \dot{x}) \quad (15)$$

Furthermore, it is clear from eq. (13) that the variable  $x$  must satisfy the periodic boundary condition

$$x(\beta) = x(0) \quad (16)$$

for eq. (14) to represent a trace (namely, the initial and the final states must be the same) and that the end point is being integrated over in the path integral in eq. (14) unlike in eq. (1). (It is important to note that the original work of Matsubara is an operator description of the imaginary time, but we will not discuss it in the present article.)

In fact, as an example, let us evaluate the canonical partition function for the bosonic oscillator using this formalism [7]. The transition amplitude is already given for zero temperature in eq. (2). Now making the identifications

$$T = -i\beta, \quad x_i = x_f = x \quad (17)$$

we obtain from eqs. (2) and (13)

$$\begin{aligned} Z(\beta) &= \int dx \left( \frac{m\omega}{2\pi \sinh \beta\omega} \right)^{\frac{1}{2}} e^{-(m\omega \tanh \frac{\beta\omega}{2}) x^2} \\ &= \left( \frac{m\omega}{2\pi \sinh \beta\omega} \right)^{\frac{1}{2}} \left( \frac{\pi}{m\omega \tanh \frac{\beta\omega}{2}} \right)^{\frac{1}{2}} \\ &= \frac{e^{\frac{\beta\omega}{2}}}{e^{\beta\omega} - 1} \end{aligned} \quad (18)$$

This is, indeed, the partition function for the bosonic oscillator as can be directly verified.

The partition function, for a fermionic system, can also be similarly given a path integral representation. However, the anti-commuting nature of the fermion variables introduces one crucial difference, namely, for a fermion theory, we have

$$Z(\beta) = \int \mathcal{D}\bar{\psi} \mathcal{D}\psi e^{-S_E[\psi, \bar{\psi}]} \quad (19)$$

with anti-periodic boundary conditions [10]

$$\psi(\beta) = -\psi(0), \quad \bar{\psi}(\beta) = -\bar{\psi}(0) \quad (20)$$

The Euclidean (imaginary time) action is again defined over a finite time interval as in eq. (15). In fact, let us calculate the canonical partition function for a fermionic oscillator, as an example, from the result in eq. (5) as well as the identifications in (20) [7]. Using

$$\psi_f = -\psi_i = -\psi, \quad \bar{\psi}_f = -\bar{\psi}_i = -\bar{\psi}$$

we obtain (remember  $T = -i\beta$ )

$$\begin{aligned} Z(\beta) &= \int d\bar{\psi} d\psi e^{\frac{\beta\omega}{2}} e^{-(1+e^{-\beta\omega})\bar{\psi}\psi} \\ &= e^{\frac{\beta\omega}{2}} (1 + e^{-\beta\omega}) = 2 \cosh \frac{\beta\omega}{2} \end{aligned} \quad (21)$$

In evaluating this, we have made use of the Berezin rules of integration [19] for Grassmann variables and we note that eq. (21), indeed, gives the correct partition function for a fermionic oscillator as can be directly calculated.

Although our discussion so far has been within the context of simple quantum mechanical systems, everything we have said can be carried over to a quantum field theory. The partition function for a quantum field theory can again be written as a path integral involving a Euclidean action as

$$Z(\beta) = \int \mathcal{D}\bar{\psi} \mathcal{D}\psi \mathcal{D}\phi e^{-S_E[\phi, \psi, \bar{\psi}]} \quad (22)$$

where the Euclidean action is defined over a finite time interval and the fields satisfy the periodicity (anti-periodicity) conditions

$$\phi(\beta, \vec{x}) = \phi(0, \vec{x}), \quad \psi(\beta, \vec{x}) = -\psi(0, \vec{x}) \quad (23)$$

and so on. The discussion is slightly more involved for gauge theories and to keep things simple, we will not discuss gauge theories.

This formulation of a field theory at finite temperature is known as the imaginary time formalism or the Matsubara formalism [1] and is the oldest formalism. There are several distinguishing features of this formalism. For example, since the time interval is finite, Fourier transformation of the time variable would involve discrete energies. In other words, the Fourier transform of the propagator, say for example, at finite temperature in the imaginary time formalism, would take the general form

$$\mathcal{G}_\beta(\tau, \vec{x}) = \frac{1}{\beta} \sum_n e^{-i\omega_n \tau} \mathcal{G}_\beta(\omega_n, \vec{x}) \quad (24)$$

where  $\omega_n = \frac{n\pi}{\beta}$  with  $n = 0, \pm 1, \pm 2, \dots$ . However, from the definition of the time ordered product

$$T_\tau(\phi(\tau)\phi^\dagger(\tau')) = \theta(\tau - \tau')\phi(\tau)\phi^\dagger(\tau') \pm \theta(\tau' - \tau)\phi^\dagger(\tau')\phi(\tau) \quad (25)$$

where we have allowed for both bosonic and fermionic fields and the KMS condition in eq. (12), it follows that, for  $\tau < 0$ ,

$$\mathcal{G}_\beta(\tau, \vec{x}) = \pm \mathcal{G}_\beta(\tau + \beta, \vec{x}) \quad (26)$$

It is important to recognize that the periodicity (anti-periodicity) of the propagator arises from the definition of the time ordered product for the bosonic (fermionic) fields and the KMS condition and is not directly connected with the periodicity (anti-periodicity) of the corresponding field variables which we have discussed earlier. This periodicity (anti-periodicity) of the propagator, on the other hand, leads to the restriction that eq. (24) holds with

$$\omega_n = \begin{cases} \frac{2n\pi}{\beta} & \text{for bosons} \\ \frac{(2n+1)\pi}{\beta} & \text{for fermions} \end{cases} \quad (27)$$

where  $n = 0, \pm 1, \dots$ . These are conventionally known as the Matsubara frequencies [22].

Given this, one can now calculate the propagators for bosonic and fermionic field theories in the Matsubara formalism and they take the forms (in the momentum space)

$$\mathcal{G}_\beta(\omega_n, \vec{k}) = \frac{1}{\omega_n^2 + \vec{k}^2 + m^2} = \frac{1}{(\frac{2n\pi}{\beta})^2 + \vec{k}^2 + m^2} \quad (28)$$

$$\mathcal{S}_\beta(\omega_n, \vec{k}) = \frac{\gamma^0 \omega_n + \vec{\gamma} \cdot \vec{k} + m}{\omega_n^2 + \vec{k}^2 + m^2} = \frac{\gamma^0 (\frac{(2n+1)\pi}{\beta}) + \vec{\gamma} \cdot \vec{k} + m}{(\frac{(2n+1)\pi}{\beta})^2 + \vec{k}^2 + m^2} \quad (29)$$

Perturbative calculations can now be developed quite analogously to the zero temperature field theory. For example, given a field theory, we can read out the vertices from the Euclidean form of the action and use the propagators of eq. (28, 29) to carry out a diagrammatic calculation which would lead to the ensemble average for a given observable. It is clear that, because the time interval is finite in this formalism, the coordinate space calculation of any diagram is cumbersome. However, much like at zero temperature, the momentum space calculation is much simpler. However, one

should keep the difference in mind, namely, that, at finite temperature, the external and the internal energies are discrete as in eq. (27). Consequently, the integration over internal energies (of zero temperature) is replaced by a sum over the internal energies. More specifically, we must use

$$\int \frac{d^4 k}{(2\pi)^4} \rightarrow \frac{1}{\beta} \sum_n \int \frac{d^3 k}{(2\pi)^3} \quad (30)$$

As an example, let us consider the self-interacting scalar theory described by

$$\mathcal{L} = \frac{1}{2} \partial_\mu \phi \partial^\mu \phi - \frac{m^2}{2} \phi^2 - \frac{\lambda}{4!} \phi^4 \quad \lambda > 0 \quad (31)$$

We note that the only one loop correction in this theory is the mass correction. Rotating to Euclidean space and using the propagator for a scalar theory as given in eq. (28) as well as (30), we obtain the one loop mass correction to be

$$\begin{aligned} \Delta m^2 &= \frac{\lambda}{2\beta} \sum_n \int \frac{d^3 k}{(2\pi)^3} \frac{1}{\left(\frac{2n\pi}{\beta}\right)^2 + \vec{k}^2 + m^2} \\ &= \frac{\lambda}{2\beta} \left(\frac{\beta}{2\pi}\right)^2 \sum_n \int \frac{d^3 k}{(2\pi)^3} \frac{1}{n^2 + \left(\frac{\beta\omega_k}{2\pi}\right)^2} \end{aligned} \quad (32)$$

Here, we have introduced the notation,

$$\omega_k = (\vec{k}^2 + m^2)^{\frac{1}{2}} \quad (33)$$

The sum, in eq. (32), can be easily evaluated using the method of residues leading to

$$\sum_{n=-\infty}^{\infty} \frac{1}{n^2 + y^2} = \frac{\pi}{y} \coth \pi y \quad \text{for } y > 0 \quad (34)$$

Using this, the one loop mass correction can be determined to be [23]

$$\begin{aligned} \Delta m^2 &= \frac{\lambda}{4} \int \frac{d^3 k}{(2\pi)^3} \frac{1}{\omega_k} \coth \left( \frac{\beta\omega_k}{2} \right) \\ &= \frac{\lambda}{4} \int \frac{d^3 k}{(2\pi)^3} \frac{1}{\omega_k} + \frac{\lambda}{2} \int \frac{d^3 k}{(2\pi)^3} \frac{1}{\omega_k} \frac{1}{e^{\beta\omega_k} - 1} \\ &= \Delta m_0^2 + \Delta m_\beta^2 \end{aligned} \quad (35)$$

There are several things to note from this calculation. First, the mass correction separates into two parts – one independent of temperature and the other genuinely a finite temperature correction. The temperature independent part (zero temperature part) is divergent as is expected at zero temperature and the divergence has to be handled by the usual process of renormalization. However, the finite temperature part is completely free from ultraviolet divergence. This is a general feature of finite temperature field theories that temperature does not introduce any new ultraviolet divergence. We will return to this question later within the context of real time formalisms for finite temperature field theories. Let us also note that (see (35)) the finite temperature integrals are, in general hard to evaluate and cannot be evaluated in a closed form. However, we can always make a high temperature expansion (small  $\beta$ ) which would give the temperature dependent correction to the mass as

$$\Delta m_\beta^2 \approx \frac{\lambda}{24\beta^2} = \frac{\lambda T^2}{24} \quad (36)$$

This shows that temperature induces a mass correction which is positive. Intuitively, it is clear that this is the behavior we would expect from a particle moving in a medium and, furthermore, the positivity of this correction is crucial in the study of symmetry restoration in field theories with spontaneous symmetry breaking.



This gives a flavor of calculations at finite temperature, particularly, in the imaginary time (Matsubara) formalism. It is worth noting here that, by construction, the imaginary time formalism would describe physical systems in equilibrium quite well. Since we have traded the time variable for temperature, it is well suited to calculate static, equilibrium quantities. Slow temperature dependence can, however, be brought in by analytically rotating the final result to Minkowski time [24]. This rotation is, on the other hand, nontrivial since we only have information about quantities at discrete energy values in the Euclidean space. The imaginary time formalism is not suitable to discuss non-equilibrium phenomena.

## 4 Thermo Field Dynamics

As we have seen, in the imaginary time formalism, the time variable is traded for the temperature. However, in studying various processes, it is desirable to have the time coordinate in addition to the temperature. Formalisms where this can be achieved are known as the real time formalisms and there are two distinct, but equivalent such formalisms. In this section, we will discuss the formalism of thermo field dynamics [3, 11, 25] returning to the alternate formalism in the next section.

Let us recall from (11) that the ensemble average of any observable is given by

$$\begin{aligned}\langle \mathcal{O} \rangle_\beta &= \frac{1}{Z(\beta)} \text{Tr} e^{-\beta \mathcal{H}} \mathcal{O} \\ &= \frac{1}{Z(\beta)} \sum_n e^{-\beta E_n} \langle n | \mathcal{O} | n \rangle\end{aligned}\quad (37)$$

Here, we have assumed that the eigenvalues of  $\mathcal{H}$  are discrete, for simplicity, and that

$$\begin{aligned}\mathcal{H} |n\rangle &= E_n |n\rangle \\ \langle m | n \rangle &= \delta_{mn} \\ \sum_n |n\rangle \langle n| &= I\end{aligned}\quad (38)$$

At zero temperature, we know that the Feynman diagrams correspond to vacuum expectation values of time ordered products. Thus, intuitively, it is clear that if we can express the ensemble averages as expectation values in some vacuum (say, a thermal vacuum), then, we can take over all the diagrammatic machinery of the zero temperature field theory. The question, therefore, is whether we can define a vacuum, say  $|0, \beta\rangle$ , such that we can write any ensemble average as

$$\langle \mathcal{O} \rangle_\beta = \langle 0, \beta | \mathcal{O} | 0, \beta \rangle = \frac{1}{Z(\beta)} \sum_n e^{-\beta E_n} \langle n | \mathcal{O} | n \rangle \quad (39)$$

Let us suppose that we can define such a thermal vacuum state as a linear superposition of the states in our physical Hilbert space, namely,

$$|0, \beta\rangle = \sum_n |n\rangle \langle n | 0, \beta \rangle = \sum_n f_n(\beta) |n\rangle \quad (40)$$

This would lead to

$$\langle 0, \beta | \mathcal{O} | 0, \beta \rangle = \sum_{n,m} f_n^*(\beta) f_m(\beta) \langle n | \mathcal{O} | m \rangle \quad (41)$$

Consequently, this would coincide with eq. (39) only if

$$f_n^*(\beta) f_m(\beta) = \frac{1}{Z(\beta)} e^{-\beta E_n} \delta_{mn} \quad (42)$$

Since  $f_n$ 's are ordinary numbers and eq. (42) is more like an orthonormality condition, it is clear that we cannot satisfy this condition (and, therefore, define a thermal vacuum with the right properties) if we restrict ourselves to the original Hilbert space.

On the other hand, it is also clear from this analysis that if  $f_n$ 's, somehow, behave like a state vector, then, the condition in eq. (42) can be easily satisfied. In fact, let us introduce a fictitious system identical to our original system and denote it by a tilde system. The states in the combined Hilbert space of this doubled system would have the form

$$|n, \tilde{n}\rangle = |n\rangle \otimes |\tilde{n}\rangle$$

Let us assume that the thermal vacuum can be written as a linear superposition of states in this doubled Hilbert space of the form

$$|0, \beta\rangle = \sum_n f_n(\beta) |n, \tilde{n}\rangle = \sum_n f_n(\beta) |n\rangle \otimes |\tilde{n}\rangle \quad (43)$$

This would lead to

$$\begin{aligned} \langle 0, \beta | \mathcal{O} | 0, \beta \rangle &= \sum_{n, m} f_n^*(\beta) f_m(\beta) \langle n, \tilde{n} | \mathcal{O} | m, \tilde{m} \rangle \\ &= \sum_{n, m} f_n^*(\beta) f_m(\beta) \langle n | \mathcal{O} | m \rangle \delta_{n, m} \\ &= \sum_n f_n^*(\beta) f_n(\beta) \langle n | \mathcal{O} | n \rangle \end{aligned} \quad (44)$$

In deriving this result, we have used the fact that an operator of the original system does not act on states of the tilde system and *vice versa*. The result in eq. (44) is quite interesting because it says that if we choose

$$f_n^*(\beta) f_n(\beta) = \frac{e^{-\beta E_n}}{Z(\beta)} \quad \text{or, } f_n(\beta) = f_n^*(\beta) = \frac{e^{-\beta E_n/2}}{Z^{1/2}(\beta)} \quad (45)$$

then, eq. (44) would, indeed, coincide with the ensemble average in eq. (39).

This analysis shows that it is possible to introduce a thermal vacuum such that the ensemble average of any operator can be written as the expectation value of the operator in the thermal vacuum. The price one has to pay is that the Hilbert space needs to be doubled. The advantage, on the other hand, lies in the fact that the description would now involve both time and temperature (since we have not traded time for temperature) and all the diagrammatic methods of zero temperature field theory can now be taken over directly.

## Fermionic Oscillator

To get a flavor for things in this formalism, let us analyze in some detail the simple quantum mechanical system of the fermionic oscillator. The Hamiltonian for the system is given by ( $\hbar = 1$ )

$$H = \omega a^\dagger a \quad (46)$$

Here, the fermionic creation and annihilation operators satisfy the canonical anti-commutation relations

$$\begin{aligned} [a, a^\dagger]_+ &= 1 \\ [a, a]_+ &= [a^\dagger, a^\dagger]_+ = 0 \end{aligned} \quad (47)$$

In this case, the spectrum of the Hamiltonian is quite simple and the Hilbert space is two dimensional with the basis states given by  $|0\rangle$  and  $|1\rangle = a^\dagger|0\rangle$ .

According to the general philosophy of thermo field dynamics, we are supposed to introduce a fictitious tilde system which is identical to our original system. Thus, we define

$$\tilde{H} = \omega \tilde{a}^\dagger \tilde{a} \quad (48)$$

with the anti-commutation relations

$$\begin{aligned} [\tilde{a}, \tilde{a}^\dagger]_+ &= 1 \\ [\tilde{a}, a]_+ &= [\tilde{a}^\dagger, a^\dagger]_+ = 0 \end{aligned} \quad (49)$$

Furthermore, we assume that the creation and the annihilation operators for the tilde and the non-tilde systems anti-commute.

The Hilbert space for the combined space is now four dimensional and, following our earlier discussion, we choose the thermal vacuum to be

$$|0, \beta\rangle = f_0(\beta)|0\rangle \otimes |\tilde{0}\rangle + f_1(\beta)|1\rangle \otimes |\tilde{1}\rangle \quad (50)$$

The normalization of the thermal vacuum gives

$$\langle 0, \beta | 0, \beta \rangle = |f_0(\beta)|^2 + |f_1(\beta)|^2 = 1 \quad (51)$$

while the expectation value of the number operator gives

$$\langle 0, \beta | N | 0, \beta \rangle = \langle 0, \beta | a^\dagger a | 0, \beta \rangle = |f_1(\beta)|^2 = \frac{1}{e^{\beta\omega} + 1} \quad (52)$$

From these, we can obtain

$$f_0(\beta) = \frac{1}{\sqrt{1 + e^{-\beta\omega}}}, \quad f_1(\beta) = \frac{e^{-\beta\omega/2}}{\sqrt{1 + e^{-\beta\omega}}} \quad (53)$$

so that we can write

$$|0, \beta\rangle = \frac{1}{\sqrt{1 + e^{-\beta\omega}}} \left( |0, \tilde{0}\rangle + e^{-\beta\omega/2} |1, \tilde{1}\rangle \right) \quad (54)$$

To further understand the properties of this system, let us note that we can define a Hermitian operator in this doubled space

$$G(\theta) = -i\theta(\beta) (\tilde{a}a - a^\dagger \tilde{a}^\dagger) \quad (55)$$

This would, in turn, lead to a formally unitary operator

$$U(\beta) = e^{-iG(\theta)} \quad (56)$$

which would connect the thermal vacuum to the vacuum of the doubled space, namely,

$$U(\beta)|0, \tilde{0}\rangle = \cos \theta(\beta)|0, \tilde{0}\rangle + \sin \theta(\beta)|1, \tilde{1}\rangle = |0, \beta\rangle \quad (57)$$

provided

$$\cos \theta(\beta) = f_0(\beta) = \frac{1}{\sqrt{1 + e^{-\beta\omega}}}, \quad \sin \theta(\beta) = f_1(\beta) = \frac{e^{-\beta\omega/2}}{\sqrt{1 + e^{-\beta\omega}}} \quad (58)$$

The unitary operator would also induce a transformation on the operators of the form

$$\mathcal{O}(\beta) = U(\beta)\mathcal{O}U^\dagger(\beta) \quad (59)$$

In particular, this would give

$$\begin{aligned} a(\beta) &= \cos \theta(\beta) a - \sin \theta(\beta) \tilde{a}^\dagger \\ \tilde{a}(\beta) &= \cos \theta(\beta) \tilde{a} + \sin \theta(\beta) a^\dagger \end{aligned} \quad (60)$$

as well as their Hermitian conjugates. These operators would satisfy the same anti-commutation relations as the original ones and we can think of them as the thermal creation and annihilation operators. Consequently, we can build up the thermal Hilbert space starting from  $|0, \beta\rangle$  and the thermal creation operators.

In particular, it is trivial to check, using (57) that the thermal vacuum satisfies

$$\begin{aligned} a(\beta)|0, \beta\rangle &= (\cos \theta(\beta) a - \sin \theta(\beta) \tilde{a}^\dagger)|0, \beta\rangle = 0 \\ \tilde{a}(\beta)|0, \beta\rangle &= (\cos \theta(\beta) \tilde{a} + \sin \theta(\beta) a^\dagger)|0, \beta\rangle = 0 \end{aligned} \quad (61)$$

This is quite interesting for it says that annihilating a particle in the thermal vacuum is equivalent to creating a tilde particle and *vice versa*. Consequently, we can intuitively think of the tilde particles as kind of hole states of the particles or particle states of the heat bath. This gives a nice intuitive meaning to the doubling of the degrees of freedom in thermo field dynamics. Namely, an isolated system in thermal equilibrium really consists of two components – the original system and the heat bath.

We also note here that although the operator connecting the thermal vacuum to the vacuum in the doubled space is formally unitary, it is more like a Bogoliubov transformation. In more complicated models with an infinite number of degrees of freedom (namely, in field theories) such an operator takes us to a unitarily inequivalent Hilbert space. Let us also note here, for future use, the simple formula following from eq. (60) that

$$\begin{pmatrix} a(\beta) \\ \tilde{a}^\dagger(\beta) \end{pmatrix} = \tilde{U}(\beta) \begin{pmatrix} a \\ \tilde{a}^\dagger \end{pmatrix} \quad (62)$$

where

$$\tilde{U}(\beta) = \begin{pmatrix} \cos \theta(\beta) & -\sin \theta(\beta) \\ \sin \theta(\beta) & \cos \theta(\beta) \end{pmatrix} \quad (63)$$

Finally, let us conclude the discussion of this example by noting that the states in the thermal Hilbert space are eigenstates of neither  $H$  nor  $\tilde{H}$ . Rather they are the eigenstates of the operator

$$\hat{H} = H - \tilde{H} \quad (64)$$

Furthermore, this combination of the Hamiltonians is also invariant under the unitary transformation of (57). This is, indeed, the Hamiltonian that governs the dynamics of the combined system.

## Bosonic Oscillator

The analysis for the case of the bosonic oscillator is quite analogous to the discussion of the fermionic oscillator. Therefore, without going into too much detail, let us summarize the results. First, the Hamiltonian for the system is given by

$$H = \omega a^\dagger a \quad (65)$$

much like the fermionic oscillator. However, the creation and annihilation operators satisfy canonical commutation relations of the form

$$\begin{aligned} [a, a^\dagger] &= 1 \\ [a, a] &= [a^\dagger, a^\dagger] = 0 \end{aligned} \quad (66)$$

The Hilbert space for the bosonic oscillator is infinite dimensional with the energy eigenstates given by

$$H|n\rangle = n\omega |n\rangle, \quad n = 0, 1, 2, \dots \quad (67)$$

According to the general discussions of thermo field dynamics, we introduce an identical, but fictitious tilde system with the Hamiltonian

$$\tilde{H} = \omega \tilde{a}^\dagger \tilde{a} \quad (68)$$

The tilde creation and annihilation operators are expected to satisfy commutation relations analogous to (66). Furthermore, the tilde operators are supposed to commute with the original operators of the theory.

Following the discussion of the earlier section, we can determine the thermal vacuum state in this case to be

$$|0, \beta\rangle = (1 - e^{-\beta\omega})^{1/2} \sum_{n=0}^{\infty} e^{-n\beta\omega/2} |n, \tilde{n}\rangle \quad (69)$$

As in the fermionic oscillator, we can introduce the Hermitian operator

$$G(\theta) = -i\theta(\beta) (\tilde{a}a - a^\dagger \tilde{a}^\dagger) \quad (70)$$

and the unitary operator

$$U(\beta) = e^{-iG(\theta)} \quad (71)$$

Then, it is straightforward to check and see that the unitary operator connects the thermal vacuum to the vacuum of the doubled space provided

$$\cosh \theta(\beta) = \frac{1}{\sqrt{1 - e^{-\beta\omega}}} \quad \sinh \theta(\beta) = \frac{e^{-\beta\omega/2}}{\sqrt{1 - e^{-\beta\omega}}} \quad (72)$$

The unitary operator induces a transformation of the operators of the form

$$\mathcal{O}(\beta) = U(\beta) \mathcal{O} U^\dagger(\beta) \quad (73)$$

leading to

$$\begin{aligned} a(\beta) &= \cosh \theta(\beta) a - \sinh \theta(\beta) \tilde{a}^\dagger \\ \tilde{a}(\beta) &= \cosh \theta(\beta) \tilde{a} - \sinh \theta(\beta) a^\dagger \end{aligned}$$

and similarly for the Hermitian conjugates. As we have seen in the last section, these can be thought of as the creation and annihilation operators for the thermal Hilbert space. In particular, the thermal vacuum is easily seen to satisfy

$$\begin{aligned} a(\beta)|0, \beta\rangle &= (\cosh \theta(\beta) a - \sinh \theta(\beta) \tilde{a}^\dagger)|0, \beta\rangle = 0 \\ \tilde{a}(\beta)|0, \beta\rangle &= (\cosh \theta(\beta) \tilde{a} - \sinh \theta(\beta) a^\dagger)|0, \beta\rangle = 0 \end{aligned} \quad (74)$$

This, again, reinforces the intuitive picture of doubling in thermo field dynamics. Let us also note here, for future use, the simple formula following from (73)

$$\begin{pmatrix} a(\beta) \\ \tilde{a}^\dagger(\beta) \end{pmatrix} = \tilde{U}(\beta) \begin{pmatrix} a \\ \tilde{a}^\dagger \end{pmatrix} \quad (75)$$

where

$$\tilde{U}(\beta) = \begin{pmatrix} \cosh \theta(\beta) & -\sinh \theta(\beta) \\ -\sinh \theta(\beta) & \cosh \theta(\beta) \end{pmatrix} \quad (76)$$

## Field Theory

The extension of these results to a field theory is quite straightforward once we keep in mind that, at the free level, a quantum field theory is simply an infinite collection of oscillators with frequencies dependent on the momentum of the mode. Consequently, the thermal vacuum, in this case, would be connected to the vacuum of the doubled space as

$$|0, \beta\rangle = U(\beta)|0, \tilde{0}\rangle = e^{-iG(\theta)} |0, \tilde{0}\rangle \quad (77)$$

where

$$G(\theta) = -i \sum_{\vec{k}} \theta_{\vec{k}}(\beta) (\tilde{a}_{\vec{k}} a_{\vec{k}} - a_{\vec{k}}^\dagger \tilde{a}_{\vec{k}}^\dagger) \quad (78)$$

with, say, for bosons,

$$\cosh \theta_{\vec{k}}(\beta) = \frac{1}{\sqrt{1 - e^{-\beta\omega_k}}} \quad \sinh \theta_{\vec{k}}(\beta) = \frac{e^{-\beta\omega_k/2}}{\sqrt{1 - e^{-\beta\omega_k}}} \quad (79)$$

Here, for a relativistic theory, we have

$$\omega_k = \sqrt{\vec{k}^2 + m^2}$$

Let us next note that, at zero temperature, the original fields are decoupled from the tilde fields. Thus, if we were to define a doublet of fields (real scalar field) as in eq. (75)

$$\Phi = \begin{pmatrix} \phi \\ \tilde{\phi} \end{pmatrix} \quad (80)$$

then, at zero temperature the propagator is defined to be (This is not to be confused with the generator of the Bogoliubov transformations in eqs. (55), (70) and (78))

$$iG(x-y) = \langle 0, \tilde{0} | T(\Phi(x)\Phi(y)) | 0, \tilde{0} \rangle$$

which has the momentum space representation

$$G(k) = \begin{pmatrix} \frac{1}{k^2 - m^2 + i\epsilon} & 0 \\ 0 & -\frac{1}{k^2 - m^2 - i\epsilon} \end{pmatrix} \quad (81)$$

Given this, the finite temperature propagator can be determined to be

$$\begin{aligned} iG_\beta(x-y) &= \langle 0, \beta | T(\Phi(x)\Phi(y)) | 0, \beta \rangle \\ &= \langle 0, \tilde{0} | U^\dagger(\beta) T(\Phi(x)\Phi(y)) U(\beta) | 0, \tilde{0} \rangle \end{aligned} \quad (82)$$

Using now the generalization of eqs. (73,75,76), the momentum representation for the propagator can be determined to be

$$\begin{aligned} G_\beta(k) &= \tilde{U}(-\theta_{\vec{k}}) G(k) \tilde{U}^T(-\theta_{\vec{k}}) \\ &= \begin{pmatrix} \frac{1}{k^2 - m^2 + i\epsilon} & 0 \\ 0 & -\frac{1}{k^2 - m^2 - i\epsilon} \end{pmatrix} - 2i\pi n_B(|k^0|) \delta(k^2 - m^2) \begin{pmatrix} 1 & e^{\beta|k^0|/2} \\ e^{\beta|k^0|/2} & 1 \end{pmatrix} \end{aligned} \quad (83)$$

There are several things to note from the structure of the propagator which are quite general for a real time formalism. First, the propagator is a  $2 \times 2$  matrix, a consequence of the doubling of the degrees of freedom. Second, the propagator is a sum of two parts – one representing the zero temperature part and the other representing the true temperature dependent corrections. The propagator is still the Greens function for the free operator of the theory, but corresponding to different boundary conditions (remember the KMS condition in eq. (12)). While the zero temperature part of the propagator corresponds, as usual, to the exchange of a virtual particle, the temperature dependent part represents an on-shell contribution (because of the delta function). In fact, the intuitive meaning of the temperature dependent correction is quite clear. In a hot medium, there is a distribution of real particles and the temperature dependent part merely represents the possibility that a particle, in addition to having virtual exchanges, can also emit or absorb a real particle of the medium.

Since the temperature dependent part of the propagator is on-shell, it is clear that there can be no new ultraviolet divergence generated at finite temperature. All the counter terms needed to renormalize the theory at zero temperature would be sufficient for studies at finite temperature as well. (Of course, the infrared behavior is another story. Infrared divergence, in a field theory, becomes much more severe at finite temperature, a topic that I will not get into.) There is an alternate way to visualize this. At finite temperature the distribution of the real particles is

Boltzmann suppressed as we go up in energy and, consequently, thermal corrections corresponding to infinite energy cannot arise.

Once, we have the propagator, we can venture to do a diagrammatic calculation in this formalism. The only things missing are the interaction vertices of the theory. There is a well defined procedure [26] (called the tilde conjugation rule) to construct the complete Lagrangian from which to construct the vertices. Very simply, it corresponds to what we have noted earlier, namely, the dynamical Hamiltonian (and, similarly, the Lagrangian) is as given in eq. (64). It is simply the difference between the original and the tilde Hamiltonians. Thus, we see that the complete theory would contain two kinds of vertices – one for the original fields while the second for the tilde fields. The vertices for the tilde fields will have a relative negative sign corresponding to the original vertices. Given the vertices and the propagator, it is now straightforward to carry out any diagrammatic calculation to any order. Let me emphasize here that although, at the tree level, there is no vertex containing both the original and the tilde fields, such vertices would be generated at higher loops because of the nontrivial matrix structure of the propagator.

Thermo Field dynamics is a real time formalism. But, more than that, it is really an operator formalism and hence very well suited to study various operator questions such as the structure of the thermal vacuum, the theorems on symmetry breaking etc. It can also be given a path integral representation and corresponds to choosing a specific time contour in the complex  $t$  plane [27, 10] (remember that the imaginary time formalism also corresponds to choosing a specific time contour, namely, along the imaginary time axis) and I will come back to this question in the next section. However, once again from the philosophy of thermo field dynamics, it is clear that, it is a natural formalism to describe equilibrium phenomena where quantities depend on both time and temperature. While there are several attempts to generalize this to include non-equilibrium phenomena, there does not yet exist a complete description.

## 5 Closed Time Path Formalism

The closed time path formalism is also a real time formalism which was formulated much earlier than thermo field dynamics within the context of non-equilibrium phenomena [2]. The two formalisms are, in some sense, complementary to each other although the closed time path formalism can describe both equilibrium and non-equilibrium phenomena with equal ease.

The basic idea behind the closed time path formalism [10] is the fact that when a quantum mechanical system is in a mixed state, as is the case in the presence of a heat bath, the system can be naturally described in terms of a density matrix defined, in the Schrödinger picture, as

$$\rho(t) = \sum_n p_n |\psi_n(t)\rangle \langle \psi_n(t)| \quad (84)$$

Here,  $p_n$  represents the probability for finding the quantum mechanical system in the state  $|\psi_n(t)\rangle$  and, for simplicity, we have assumed the quantum mechanical states to form a discrete set. It is  $p_n$  which contains information regarding the surrounding which is hard to determine, but, being a probability, it satisfies

$$\sum_n p_n = 1$$

Given the density matrix, the ensemble average of any operator can be calculated in the Schrödinger picture as

$$\langle \mathcal{O} \rangle(t) = \sum_n p_n \langle \psi_n(t) | \mathcal{O} | \psi_n(t) \rangle = \text{Tr } \rho(t) \mathcal{O} \quad (85)$$

The ensemble average, in this case, naturally develops a time dependence from the time dependence of the density matrix. In this formalism, we can naturally define an entropy as

$$S = - \sum_n p_n \ln p_n$$

which is by definition positive semi-definite and measures the order (or lack of it) in an ensemble.

The state vectors satisfy the Schrödinger equation ( $\hbar = 1$ )

$$i \frac{\partial |\psi(t)\rangle}{\partial t} = H |\psi(t)\rangle$$

From this, we can determine the time evolution of the density matrix which turns out to be the Liouville equation

$$i \frac{\partial \rho(t)}{\partial t} = [H, \rho(t)] \quad (86)$$

In deriving this, we have assumed that the probabilities do not change with time (appreciably) implying that entropy remains constant during such an evolution. The reason for this assumption is our lack of knowledge about the time evolution of the surrounding such as the heat bath. On the other hand, adiabatic evolutions do arise frequently in physical systems and, consequently, we would continue with this assumption.

Let us note that eq. (86) has a simple solution of the form

$$\rho(t) = U(t, 0) \rho(0) U^\dagger(t, 0) = U(t, 0) \rho(0) U(0, t) \quad (87)$$

where the time evolution operator has the general form

$$U(t, t') = T \left( e^{-i \int_{t'}^t dt'' H(t'')} \right) \quad (88)$$

Furthermore, it satisfies the semi-group properties

$$\begin{aligned} U(t_1, t_2) U(t_2, t_1) &= 1 \\ U(t_1, t_2) U(t_2, t_3) &= U(t_1, t_3) \quad \text{for } t_1 > t_2 > t_3 \end{aligned} \quad (89)$$

In particular, let us note that if the Hamiltonian is time independent, eq. (87) takes the simple form

$$\rho(t) = e^{-iHt} \rho(0) e^{iHt}$$

and, furthermore, if the Hamiltonian commutes with  $\rho(0)$ , the density matrix would be time independent, describing a system in equilibrium. This would be true, for example, if the states in eq. (84) are stationary states. This is also true if the probabilities have a Boltzmann distribution in which case, we refer to the system as being in thermal equilibrium. However, we will not restrict to any such special case allowing for the formalism to accommodate both equilibrium and non-equilibrium phenomena.

Keeping in mind the fact that we are ultimately interested in a thermal ensemble, let us choose

$$\rho(0) = \frac{e^{-\beta H_i}}{\text{Tr } e^{-\beta H_i}} \quad (90)$$

for some  $H_i$ . Since the density matrix is a positive Hermitian matrix with unit trace, mathematically, this is allowed. But, more important is the physical reason behind such a choice. Namely, we can think of the dynamical Hamiltonian of our system as

$$H(t) = \begin{cases} H_i & \text{for } \text{Re } t \leq 0 \\ \mathcal{H}(t) & \text{for } \text{Re } t \geq 0 \end{cases} \quad (91)$$

This would correspond to the fact that we prepare our system in a equilibrium state at temperature  $\frac{1}{\beta}$  for negative times and let the system evolve, for positive times, with the true Hamiltonian  $\mathcal{H}$  which may be time dependent. If  $\mathcal{H}(t) = H_i$ , then, the system will evolve in equilibrium and not otherwise.

With eq. (91) in mind, we note that we can write

$$\rho(0) = \frac{U(T - i\beta, T)}{\text{Tr } U(T - i\beta, T)} \quad (92)$$



where  $T$  is assumed to be a large negative time (and not the temperature) and  $T \rightarrow -\infty$  at the end. Using the semi-group properties of the time evolution operator, it is easy to see that the ensemble average of any operator can now be represented as

$$\begin{aligned}
 \langle \mathcal{O} \rangle_\beta &= \text{Tr } \rho(t) \mathcal{O} \\
 &= \frac{\text{Tr } U(t, 0) U(T - i\beta, T) U(0, t) \mathcal{O}}{\text{Tr } U(T - i\beta, T)} \\
 &= \frac{\text{Tr } U(T - i\beta, T) U(T, T') U(T', t) \mathcal{O} U(t, T)}{\text{Tr } U(T - i\beta, T) U(T, T') U(T', T)}
 \end{aligned} \tag{93}$$

where we have introduced a large positive time  $T'$  and assume that  $T' \rightarrow \infty$  at the end. This gives a nice representation to the ensemble average of any operator. Namely, we let the system evolve from a large negative time  $T$  to  $t$  where the appropriate operator  $\mathcal{O}$  is inserted. The system then, evolves from  $t$  to a large positive time  $T'$  and back from  $T'$  to  $T$  and then, continues evolving along the imaginary branch from  $T$  to  $T - i\beta$ . Since the matrix elements of the time evolution operator can be given a path integral representation, it is clear that the ensemble average of any operator can also be given a path integral representation in this formalism corresponding to the specific contour in the complex time plane as described above. Although the specific contour has three branches – one along the real axis increasing with time, the second also along the real axis decreasing with time and the third along the negative imaginary axis – in the limit  $T \rightarrow -\infty$  and  $T' \rightarrow \infty$ , it can be shown that the third branch gets decoupled from the other two (the factors in the propagators connecting such branches are asymptotically damped). Consequently, in this limit, we are effectively dealing with two branches leading to the name “closed time path formalism” [12]. In this contour, then, the time integration has to be thought of as

$$\int_c dt = \int_{-\infty}^{\infty} dt_+ - \int_{-\infty}^{\infty} dt_- \tag{94}$$

where the relative negative sign arises because time is decreasing in the second branch of the time contour.

The doubling of the degrees of freedom, in this formalism, is now clear. To have a path integral description, we must specify the fields on both the branches of the contour. Or, equivalently, we can use just the positive branch and double the field degrees of freedom. Namely, corresponding to every original field, say  $\phi_+$ , we must introduce a second field  $\phi_-$  and remember that the action for the  $\phi_-$  fields must have a relative negative sign arising from eq. (94), namely, that time is decreasing along the second branch.

## Scalar Field Theory

Just as an example, let us study next the self-interacting scalar field theory in some detail. The Lagrangian density is the same as in eq. (30), but following the earlier discussion, we should take the complete Lagrangian density for the system to be

$$\mathcal{L} = \mathcal{L}(\phi_+) - \mathcal{L}(\phi_-) \tag{95}$$

where

$$\mathcal{L}(\phi) = \frac{1}{2} \partial_\mu \phi \partial^\mu \phi - \frac{m^2}{2} \phi^2 - \frac{\lambda}{4!} \phi^4 \quad \lambda > 0 \tag{96}$$

The Feynman propagator can again be determined for this theory and would have a  $2 \times 2$  matrix structure because of the doubling of the field degrees of freedom. It can be determined subject to compatibility with the KMS conditions and has the form in the momentum space

$$G(k) = \begin{pmatrix} G_{++}(k) & G_{+-}(k) \\ G_{-+}(k) & G_{--}(k) \end{pmatrix} \tag{97}$$

with

$$\begin{aligned}
G_{++}(k) &= \frac{1}{k^2 - m^2 + i\epsilon} - 2i\pi n_B(|k^0|) \delta(k^2 - m^2) \\
G_{+-}(k) &= -2i\pi (\theta(-k^0) + n_B(|k^0|)) \delta(k^2 - m^2) \\
G_{-+}(k) &= -2i\pi (\theta(k^0) + n_B(|k^0|)) \delta(k^2 - m^2) \\
G_{--}(k) &= \frac{1}{k^2 - m^2 - i\epsilon} - 2i\pi n_B(|k^0|) \delta(k^2 - m^2)
\end{aligned} \tag{98}$$

There are several things to note from the structure of this propagator. First, as in the case of the propagator in thermo field dynamics, here, too, we see that the propagator naturally is a sum of two parts – the temperature independent part and the temperature dependent part. But, more interestingly, here the propagator has the simplification that the temperature dependent part of every component is the same which leads to various simplifications in actual studies of thermal quantities. Furthermore, not all the components of the propagator are independent. In fact, it is easily seen that (this can be traced back to their definition)

$$G_{++}(k) + G_{--}(k) = G_{+-}(k) + G_{-+}(k)$$

These are known as the causal propagators of the theory and are useful in diagrammatic evaluation. There is, of course, another kind of propagator, conventionally known as the physical propagators and is defined as

$$\hat{G}(k) = \begin{pmatrix} 0 & G_A(k) \\ G_R(k) & G_C(k) \end{pmatrix} \tag{99}$$

where  $G_A$ ,  $G_R$  and  $G_C$  are known as the advanced, retarded and the correlated Greens functions. These are quite useful in the study of various phenomena such as the linear response theory. The important thing to observe is that the causal and the physical propagators are connected through a unitary transformation

$$\hat{G}(k) = Q G Q^\dagger \tag{100}$$

where

$$Q = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \tag{101}$$

It can be determined from this that, at the tree level,

$$\begin{aligned}
G_A(k) &= \frac{1}{k^2 - m^2 - i\epsilon k^0} \\
G_R(k) &= \frac{1}{k^2 - m^2 + i\epsilon k^0} \\
G_C(k) &= -2i\pi (1 + 2n_B(|k^0|)) \delta(k^2 - m^2)
\end{aligned} \tag{102}$$

as they should be.

The diagrammatic calculations can now be easily understood in this formalism. The vertices can be read out from the Lagrangian density in eq. (95). There are two kinds of vertices, one for the original fields,  $\phi_+$ , and the other for the doubled fields,  $\phi_-$ . The vertices for the  $\phi_-$  fields are the same as those for the  $\phi_+$  fields except for a relative sign. With the vertices and the causal propagators, one can now carry out the calculation of any observable to any order in perturbation theory. As before, we note that, although there is no coupling between the  $\phi_+$  and  $\phi_-$  fields at the tree level, higher order corrections would, in general, couple them.

As an example, let us calculate the one loop mass correction in this theory. There will be two such diagrams to calculate – one for the  $\phi_+$  field and the other for the  $\phi_-$  field. The mass correction for the  $\phi_+$  field is readily seen to be

$$-i\Delta m_+^2 = \frac{(-i\lambda)}{2} \int \frac{d^4 k}{(2\pi)^4} iG_{++}(k)$$

$$\begin{aligned}
&= \frac{\lambda}{2} \int \frac{d^4 k}{(2\pi)^4} \left( \frac{1}{k^2 - m^2 + i\epsilon} - 2i\pi n_B(|k^0|) \delta(k^2 - m^2) \right) \\
&= -i(\Delta m_0^2 + \Delta m_\beta^2)
\end{aligned} \tag{103}$$

where it is easily seen that the temperature independent part has the form

$$\Delta m_0^2 = \frac{\lambda}{4} \int \frac{d^3 k}{(2\pi)^3} \frac{1}{\omega_k} \tag{104}$$

while the temperature dependent part is given by

$$\Delta m_\beta^2 = \frac{\lambda}{2} \int \frac{d^3 k}{(2\pi)^3} \frac{n_B(\omega_k)}{\omega_k} = \frac{\lambda}{2} \int \frac{d^3 k}{(2\pi)^3} \frac{1}{\omega_k} \frac{1}{e^{\beta\omega_k} - 1} \tag{105}$$

These can be compared with the corresponding terms in eq. (35). We can also calculate the mass correction for the  $\phi_-$  field. With a little bit of analysis, it is seen that

$$\Delta m_-^2 = \Delta m_+^2$$

## 6 Feynman Parameterization

So far, we have described the various formalisms that can be used to do calculations at finite temperature. However, actual calculations lead to many subtle, but interesting features of theories at finite temperature. One immediate and obvious feature, of course, is that finite temperature effects break Lorentz invariance. Namely, in studying a system at finite temperature, one has to go to a specific frame where the heat bath is at rest and, consequently, Lorentz invariance will no longer be manifest. This is, of course, already manifest at the level of propagators. For example, the structure of the propagators in eqs. (83) or (98) clearly displays a Lorentz non-invariant structure. The consequence of this is that an amplitude calculated at finite temperature, say for example, the self-energy  $\Pi(p^0, \vec{p})$  depends on the external energy and momentum independently. In fact, the self-energy becomes a non-analytic function of these two variables at the origin and two different ways of approaching the origin in this space leads to distinct plasmon and screening masses [23]. Thus, such non-analyticities are quite physical and their origin can be traced back to the fact that, at finite temperature, there are new channels of reactions possible leading to new branch cuts which give rise to such discontinuities [23, 10]. (To be absolutely fair, it is worth noting that statistical mechanics can be formulated in a covariant way. In such a case, one finds that there is a larger number of Lorentz invariant variables that can be defined on which amplitudes can depend. The non-analyticity in  $p^0$  and  $\vec{p}$  can then be translated to a non-analyticity in these new, Lorentz invariant variables [23].)

There are, of course, some other kinds of subtlety that arise which influence the calculations directly at finite temperature. We will discuss one such subtlety in this section. Let us note that a particularly useful formula in the evaluation of amplitudes at zero temperature is the Feynman combination formula given by

$$\frac{1}{A + i\epsilon} \frac{1}{B + i\epsilon} = \int_0^1 \frac{dx}{[x(A + i\epsilon) + (1-x)(B + i\epsilon)]^2} \tag{106}$$

This can be directly checked by evaluating the  $x$  integral on the right hand side.

This formula is extremely useful and works at zero temperature mainly because the Feynman propagators have the same analytic structure, namely, they have the same “ $i\epsilon$ ” dependence. In contrast, we note that the finite temperature propagators contain delta functions (see eqs. (83) and (98)) and recalling that

$$\delta(x) = \lim_{\epsilon \rightarrow 0^+} \frac{1}{2i\pi} \left( \frac{1}{x - i\epsilon} - \frac{1}{x + i\epsilon} \right)$$

we recognize that, at finite temperature, the propagators no longer have the same “ $i\epsilon$ ” dependence. Consequently, in evaluating Feynman amplitudes at finite temperature, we have to combine denominators which do not necessarily have the same “ $i\epsilon$ ” dependence. Keeping this in mind, let us examine the combination of two different denominators with arbitrary analytic dependence.

Without loss of generality, let us choose  $\alpha, \beta = \pm 1$  and note that

$$\begin{aligned} \int_0^1 \frac{dx}{[x(A + i\alpha\epsilon) + (1-x)(B + i\beta\epsilon)]^2} &= -\frac{1}{(A-B) + i(\alpha-\beta)\epsilon} \times \\ &\quad \left[ \frac{1}{x(A + i\alpha\epsilon) + (1-x)(B + i\beta\epsilon)} \right]_0^1 \\ &= \frac{1}{A + i\alpha\epsilon} - \frac{1}{B + i\beta\epsilon} \end{aligned} \quad (107)$$

This is, of course, the standard Feynman combination formula. However, let us note that this will not hold if  $0 < x_0 < 1$  such that

$$x_0 = \frac{\beta}{\beta - \alpha}, \quad \beta A = \alpha B \quad (108)$$

because, in such a case, the integrand will have a singularity on the real  $x$ -axis inside the interval of integration. In this case, we have

$$\begin{aligned} \int_0^1 \frac{dx}{[x(A + i\alpha\epsilon) + (1-x)(B + i\beta\epsilon)]^2} &= \int_0^1 \frac{dx}{[x(A - B + i(\alpha - \beta)\epsilon) + B + i\beta\epsilon]^2} \\ &= \lim_{\eta \rightarrow 0} \left( \int_0^{\frac{\beta}{\beta - \alpha} - \eta} + \int_{\frac{\beta}{\beta - \alpha} + \eta}^1 \right) \times \\ &\quad \frac{dx}{[x(A - B + i(\alpha - \beta)\epsilon) + B + i\beta\epsilon]^2} \\ &= \frac{1}{A + i\alpha\epsilon} - \frac{1}{B + i\beta\epsilon} \\ &\quad - 2i\pi \frac{(\alpha - \beta)\delta(\beta A - \alpha B)}{A - B + i(\alpha - \beta)\epsilon} \end{aligned} \quad (109)$$

In other words, when the parameters of the integrand satisfy eq. (108), the Feynman combination formula of eq. (106) will modify and the general formula follows from eq. (109) to be [13]

$$\frac{1}{A + i\alpha\epsilon} - \frac{1}{B + i\beta\epsilon} = \int_0^1 \frac{dx}{[x(A + i\alpha\epsilon) + (1-x)(B + i\beta\epsilon)]^2} + 2i\pi \frac{(\alpha - \beta)\delta(\beta A - \alpha B)}{A - B + i(\alpha - \beta)\epsilon} \quad (110)$$

Note that, condition (108) can only be satisfied (with  $0 < x_0 < 1$ ) if  $\alpha$  and  $\beta$  are of opposite sign. Indeed, let us note from eq. (110) that the second term vanishes when  $\alpha = \beta = 1$  as is the case at zero temperature. Namely, when propagators with identical “ $i\epsilon$ ” dependence are combined, the standard combination formula of eq. (106) holds. However, if denominators with opposite “ $i\epsilon$ ” dependence are combined, the correct combination formula involves a second term. This is quite crucial at finite temperature and without this second term, one ends up with a wrong result as was discovered in finite temperature calculations the hard way [28].

## 7 Large Gauge Invariance

Gauge theories are beautiful theories which describe physical forces in a natural manner and because of their rich structure, the study of gauge theories at finite temperature is quite interesting in itself. However, to avoid getting into technicalities, we will not discuss the intricacies of such theories either at zero temperature or at finite temperature. Rather, we will discuss a simple quantum

mechanical model, in this section, to bring out some of the new features that temperature brings into such theories – features which are very different from what we expect at zero temperature.

To motivate, let us note that gauge invariance is realized as an internal symmetry in quantum mechanical systems. Consequently, we do not expect a macroscopic external surrounding such as a heat bath to modify gauge invariance. This is more or less what is also found by explicit computations at finite temperature, namely, that gauge invariance and Ward identities continue to hold even at finite temperature [29]. This is certainly the case when one is talking about small gauge transformations for which the parameters of transformation vanish at infinity.

However, there is a second class of gauge invariance, commonly known as large gauge invariance where the parameters do not vanish at infinity and this brings in some new topological character to physical theories. For example, let us consider a 2 + 1 dimensional Chern-Simons theory of the form

$$\begin{aligned}\mathcal{L} &= M\mathcal{L}_{\text{CS}} + \mathcal{L}_{\text{fermion}} \\ &= M\epsilon^{\mu\nu\lambda} \text{tr} A_\mu(\partial_\nu A_\lambda - \frac{2}{3}A_\nu A_\lambda) + \bar{\psi}(\gamma^\mu(i\partial_\mu - gA_\mu) - m)\psi\end{aligned}\quad (111)$$

where  $M$  is a mass parameter,  $A_\mu$  a matrix valued non-Abelian gauge field and “tr” stands for the matrix trace. The first term, on the right hand side, is known as the Chern-Simons term which exists only in odd space-time dimensions. We can, of course, also add a Maxwell like term to the Lagrangian and, in that case, the Chern-Simons term behaves like a mass term for the gauge field. Consequently, such a term is also known as a topological mass term [30] (topological because it does not involve the metric). For simplicity of discussion, however, we will not include a Maxwell like term to the Lagrangian.

Under a gauge transformation of the form

$$\begin{aligned}\psi &\rightarrow U^{-1}\psi \\ A_\mu &\rightarrow U^{-1}A_\mu U - \frac{i}{g}U^{-1}\partial_\mu U\end{aligned}\quad (112)$$

it is straightforward to check that the action in eq. (111) is not invariant, rather it changes as

$$S = \int d^3x \mathcal{L} \rightarrow S + \frac{4\pi M}{g^2} 2i\pi W \quad (113)$$

where

$$W = \frac{1}{24\pi^2} \int d^3x \epsilon^{\mu\nu\lambda} \text{tr} \partial_\mu U U^{-1} \partial_\nu U U^{-1} \partial_\lambda U U^{-1} \quad (114)$$

is known as the winding number. It is a topological quantity which is an integer (Basically, the fermion Lagrangian density is invariant under the gauge transformations, but the Chern-Simons term changes by a total divergence which does not vanish if the gauge transformations do not vanish at infinity. Consequently, the winding number counts the number of times the gauge transformations wrap around the sphere.). For small gauge transformations, the winding number vanishes since the gauge transformations vanish at infinity.

Let us note from eq. (113) that even though the action is not invariant under a large gauge transformation, if  $M$  is quantized in units of  $\frac{g^2}{4\pi^2}$ , the change in the action would be a multiple of  $2i\pi$  and, consequently, the path integral would be invariant under a large gauge transformation. Thus, we have the constraint coming from the consistency of the theory that the coefficient of the Chern-Simons term must be quantized. We have derived this conclusion from an analysis of the tree level behavior of the theory and we have to worry if the quantum corrections can change the behavior of the theory. At zero temperature, an analysis of the quantum corrections shows that the theory continues to be well defined with the tree level quantization of the Chern-Simons coefficient provided the number of fermion flavors is even. The even number of fermion flavors is also necessary for a global anomaly of the theory to vanish and so, everything is well understood at zero temperature.

At finite temperature, however, the situation appears to change drastically. Namely, the fermions induce a temperature dependent Chern-Simons term effectively making [31]

$$M \rightarrow M - \frac{g^2}{4\pi} \frac{mN_f}{2|m|} \tanh \frac{\beta|m|}{2} \quad (115)$$

Here,  $N_f$  is the number of fermion flavors and this shows that, at zero temperature ( $\beta \rightarrow \infty$ ),  $M$  changes by an integer (in units of  $g^2/4\pi$ ) for an even number of flavors. However, at finite temperature, this becomes a continuous function of temperature and, consequently, it is clear that it can no longer be an integer for arbitrary values of the temperature. It seem, therefore, that temperature would lead to a breaking of large gauge invariance in such a system. This is, on the other hand, completely counter intuitive considering that temperature should have no direct influence on gauge invariance of the theory.

### C-S Theory in 0 + 1 Dimension

As we have noted, Chern-Simons terms can exist in odd space-time dimensions. Consequently, let us try to understand this puzzle of large gauge invariance in a simple quantum mechanical theory. Let us consider a simple theory of an interacting massive fermion with a Chern-Simons term in 0 + 1 dimension described by [14, 32]

$$L = \bar{\psi}_j(i\partial_t - A - m)\psi_j - \kappa A \quad (116)$$

Here,  $j = 1, 2, \dots, N_f$  labels the fermion flavors. There are several things to note from this. First, we are considering an Abelian gauge field for simplicity. Second, in this simple model, the gauge field has no dynamics (in 0 + 1 dimension the field strength is zero) and, therefore, we do not have to get into the intricacies of gauge theories. There is no Dirac matrix in 0 + 1 dimension as well making the fermion part of the theory quite simple as well. And, finally, the Chern-Simons term, in this case, is a linear field so that we can, in fact, think of the gauge field as an auxiliary field.

In spite of the simplicity of this theory, it displays a rich structure including all the properties of the 2 + 1 dimensional theory that we have discussed earlier. For example, let us note that under a gauge transformation

$$\psi_j \rightarrow e^{-i\lambda(t)}\psi_j, \quad A \rightarrow A + \partial_t\lambda(t) \quad (117)$$

the fermion part of the Lagrangian is invariant, but the Chern-Simons term changes by a total derivative giving

$$S = \int dt L \rightarrow S - 2\pi\kappa N \quad (118)$$

where

$$N = \frac{1}{2\pi} \int dt \partial_t \lambda(t) \quad (119)$$

is the winding number and is an integer which vanishes for small gauge transformations. Let us note that a large gauge transformation can have a parametric form of the form, say,

$$\lambda(t) = -iN \log \left( \frac{1+it}{1-it} \right) \quad (120)$$

The fact that  $N$  has to be an integer can be easily seen to arise from the requirement of single-valuedness for the fermion field. Once again, in light of our earlier discussion, it is clear from eq. (118) that the theory is meaningful only if  $\kappa$ , the coefficient of the Chern-Simons term, is an integer.

Let us assume, for simplicity, that  $m > 0$  and compute the correction to the photon one-point function arising from the fermion loop at zero temperature.

$$iI_1 = -(-i)N_f \int \frac{dk}{2\pi} \frac{i(k+m)}{k^2 - m^2 + i\epsilon} = \frac{iN_f}{2} \quad (121)$$

This shows that, as a result of the quantum correction, the coefficient of the Chern-Simons term would change as

$$\kappa \rightarrow \kappa - \frac{N_f}{2}$$

As in 2 + 1 dimensions, it is clear that the coefficient of the Chern-Simons term would continue to be quantized and large gauge invariance would hold if the number of fermion flavors is even. At zero temperature, we can also calculate the higher point functions due to the fermions in the theory and they all vanish. This has a simple explanation following from the small gauge invariance of the theory. Namely, suppose we had a nonzero two point function, then, it would imply a quadratic term in the effective action of the form

$$\Gamma_2 = \frac{1}{2} \int dt_1 dt_2 A(t_1) F(t_1 - t_2) A(t_2) \quad (122)$$

Furthermore, invariance under a small gauge transformation would imply

$$\delta\Gamma_2 = - \int dt_1 dt_2 \lambda(t_1) \partial_{t_1} F(t_1 - t_2) A(t_2) = 0 \quad (123)$$

The solution to this equation is that  $F = 0$  so that there cannot be a quadratic term in the effective action which would be local and yet be invariant under small gauge transformations. A similar analysis would show that small gauge invariance does not allow any higher point function to exist at zero temperature.

Let us also note that eq. (123) has another solution, namely,

$$F(t_1 - t_2) = \text{constant}$$

In such a case, however, the quadratic action becomes non-extensive, namely, it is the square of an action. We do not expect such terms to arise at zero temperature and hence the constant has to vanish for vanishing temperature. As we will see next, the constant does not have to vanish at finite temperature and we can have non-vanishing higher point functions implying a non-extensive structure of the effective action.

The fermion propagator at finite temperature (in the real time formalism) has the form [10]

$$\begin{aligned} S(p) &= (p + m) \left( \frac{i}{p^2 - m^2 + i\epsilon} - 2\pi n_F(|p|) \delta(p^2 - m^2) \right) \\ &= \frac{i}{p - m + i\epsilon} - 2\pi n_F(m) \delta(p - m) \end{aligned} \quad (124)$$

and the structure of the effective action can be studied in the momentum space in a straightforward manner. However, in this simple model, it is much easier to analyze the amplitudes in the coordinate space. Let us note that the coordinate space structure of the fermion propagator is quite simple, namely,

$$S(t) = \int \frac{dp}{2\pi} e^{-ipt} \left( \frac{i}{p - m + i\epsilon} - 2\pi n_F(m) \delta(p - m) \right) = (\theta(t) - n_F(m)) e^{-imt} \quad (125)$$

In fact, the calculation of the one point function is trivial now

$$iI_1 = -(-i)N_f S(0) = \frac{iN_f}{2} \tanh \frac{\beta m}{2} \quad (126)$$

This shows that the behavior of this theory is completely parallel to the 2 + 1 dimensional theory in that, it would suggest

$$\kappa \rightarrow \kappa - \frac{N_f}{2} \tanh \frac{\beta m}{2}$$

and it would appear that large gauge invariance would not hold at finite temperature.

Let us next calculate the two point function at finite temperature.

$$\begin{aligned}
 iI_2 &= -(-i)^2 \frac{N_f}{2!} S(t_1 - t_2) S(t_2 - t_1) \\
 &= -\frac{N_f}{2} n_F(m) (1 - n_F(m)) \\
 &= -\frac{N_f}{8} \operatorname{sech}^2 \frac{\beta m}{2} = \frac{1}{2} \frac{1}{2!} \frac{i}{\beta} \frac{\partial(iI_1)}{\partial m}
 \end{aligned} \tag{127}$$

This shows that the two point function is a constant as we had noted earlier implying that the quadratic term in the effective action would be non-extensive.

Similarly, we can also calculate the three point function trivially and it has the form

$$iI_3 = \frac{iN_f}{24} \tanh \frac{\beta m}{2} \operatorname{sech}^2 \frac{\beta m}{2} = \frac{1}{2} \frac{1}{3!} \left( \frac{i}{\beta} \right)^2 \frac{\partial^2(iI_1)}{\partial m^2} \tag{128}$$

In fact, all the higher point functions can be worked out in a systematic manner. But, let us observe a simple method of computation for these. We note that because of the gauge invariance (Ward identity), the amplitudes cannot depend on the external time coordinates as is clear from the calculations of the lower point functions. Therefore, we can always simplify the calculation by choosing a particular time ordering convenient to us. Second, since we are evaluating a loop diagram (a fermion loop) the initial and the final time coordinates are the same and, consequently, the phase factors in the propagator (125) drop out. Therefore, let us define a simplified propagator without the phase factor as

$$\tilde{S}(t) = \theta(t) - n_F(m) \tag{129}$$

so that we have

$$\tilde{S}(t > 0) = 1 - n_F(m), \quad \tilde{S}(t < 0) = -n_F(m) \tag{130}$$

Then, it is clear that with the choice of the time ordering,  $t_1 > t_2$ , we can write

$$\begin{aligned}
 \frac{\partial \tilde{S}(t_1 - t_2)}{\partial m} &= -\beta \tilde{S}(t_1 - t_3) \tilde{S}(t_3 - t_2) & t_1 > t_2 > t_3 \\
 \frac{\partial \tilde{S}(t_2 - t_1)}{\partial m} &= -\beta \tilde{S}(t_2 - t_3) \tilde{S}(t_3 - t_1) & t_1 > t_2 > t_3
 \end{aligned} \tag{131}$$

In other words, this shows that differentiation of a fermionic propagator with respect to the mass of the fermion is equivalent to introducing an external photon vertex (and, therefore, another fermion propagator as well) up to constants. This is the analogue of the Ward identity in QED in four dimensions except that it is much simpler. From this relation, it is clear that if we take a  $n$ -point function and differentiate this with respect to the fermion mass, then, that is equivalent to adding another external photon vertex in all possible positions. Namely, it should give us the  $(n+1)$ -point function up to constants. Working out the details, we have,

$$\frac{\partial I_n}{\partial m} = -i\beta(n+1)I_{n+1} \tag{132}$$

Therefore, the  $(n+1)$ -point function is related to the  $n$ -point function recursively and, consequently, all the amplitudes are related to the one point function which we have already calculated. (Incidentally, this is already reflected in eqs. (127,128)).

With this, we can now determine the full effective action of the theory at finite temperature to be

$$\begin{aligned}
 \Gamma &= -i \sum_n a^n (iI_n) \\
 &= -\frac{i\beta N_f}{2} \sum_n \frac{(ia/\beta)^n}{n!} \left( \frac{\partial}{\partial m} \right)^{n-1} \tanh \frac{\beta m}{2} \\
 &= -iN_f \log \left( \cos \frac{a}{2} + i \tanh \frac{\beta m}{2} \sin \frac{a}{2} \right)
 \end{aligned} \tag{133}$$



where we have defined

$$a = \int dt A(t) \quad (134)$$

There are several things to note from this result. First of all, the higher point functions are no longer vanishing at finite temperature and give rise to a non-extensive structure of the effective action. More importantly, when we include all the higher point functions, the complete effective action is invariant under large gauge transformations, namely, under

$$a \rightarrow a + 2\pi N \quad (135)$$

the effective action changes as

$$\Gamma \rightarrow \Gamma + NN_f \pi \quad (136)$$

which leaves the path integral invariant for an even number of fermion flavors. This clarifies the puzzle of large gauge invariance at finite temperature in this model. Namely, when we are talking about large changes (large gauge transformations), we cannot ignore higher order terms if they exist. This may provide a resolution to the large gauge invariance puzzle in the  $2 + 1$  dimensional theory as well. However, in spite of several nice analysis [33], this puzzle has not yet been settled in all its generality in the  $2 + 1$  dimensional case.

## Exact Result

In the earlier section, we discussed a perturbative method of calculating the effective action at finite temperature which clarified the puzzle of large gauge invariance. However, this quantum mechanical model is simple enough that we can also evaluate the effective action directly and, therefore, it is worth asking how the perturbative calculations compare with the exact result.

The exact evaluation of the effective action can be done easily using the imaginary time formalism. But, first, let us note that the fermionic part of the Lagrangian in eq. (116) has the form

$$L_f = \bar{\psi}(i\partial_t - A - m)\psi \quad (137)$$

where we have suppressed the fermion flavor index for simplicity. Let us note that if we make a field redefinition of the form

$$\psi(t) = e^{-i \int_0^t dt' A(t')} \tilde{\psi}(t) \quad (138)$$

then, the fermionic part of the Lagrangian becomes free, namely,

$$L_f = \bar{\tilde{\psi}}(i\partial_t - m)\tilde{\psi} \quad (139)$$

This is a free theory and, therefore, the path integral can be easily evaluated. However, we have to remember that the field redefinition in (138) changes the periodicity condition for the fermion fields. Since the original fermion field was expected to satisfy anti-periodicity

$$\psi(\beta) = -\psi(0)$$

it follows now that the new fields must satisfy

$$\tilde{\psi}(\beta) = -e^{-ia} \tilde{\psi}(0) \quad (140)$$

Consequently, the path integral for the free theory (139) has to be evaluated subject to the periodicity condition of (140).

Although the periodicity condition (140) appears to be complicated, it is well known that the effect can be absorbed by introducing a chemical potential [10], in the present case, of the form

$$\mu = \frac{ia}{\beta} \quad (141)$$

With the addition of this chemical potential, the path integral can be evaluated subject to the usual anti-periodicity condition. The effective action can now be easily determined

$$\begin{aligned}
 \Gamma &= -i \log \left( \frac{\det(i\partial_t - m + \frac{ia}{\beta})}{(i\partial_t - m)} \right)^{N_f} \\
 &= -iN_f \log \left( \frac{\cosh \frac{\beta}{2}(m - \frac{ia}{\beta})}{\cosh \frac{\beta m}{2}} \right) \\
 &= -iN_f \log \left( \cos \frac{a}{2} + i \tanh \frac{\beta m}{2} \sin \frac{a}{2} \right)
 \end{aligned} \tag{142}$$

which coincides with the perturbative result of eq. (134).

## 8 Supersymmetry Breaking

One of the reasons for studying finite temperature field theory is to understand questions such as phase transitions in such systems. It is by now well understood that most field theoretic models of spontaneous symmetry breaking display a phase structure much like what one sees in a magnet, namely, above a certain critical temperature, the system is in a symmetric phase while below the critical temperature, the system is in a broken symmetry phase. Thus, temperature has the almost universal effect that if a symmetry is spontaneously broken at low temperature, it is restored at temperatures above a certain critical value. Qualitatively, it can be understood as follows. Temperature, particularly high temperature, provides a lot of thermal energy to a physical system to wash out any structure in the zero temperature potential which may be responsible for symmetry breaking. There is, however, one class of symmetries where temperature has the inverse effect, namely, in a supersymmetric theory, a symmetric phase at low temperature goes to a broken phase at high temperature. (Of course, if supersymmetry is broken at low temperature, it continues to be broken even at high temperature.) We will discuss this phenomenon with a simple quantum mechanical model in this section.

### Supersymmetric Oscillator at $T = 0$

Let us note that supersymmetry is an ultimate form of symmetry that one can dream of, namely, it transforms bosons into fermions and *vice versa* [34-35]. To introduce supersymmetry, let us consider a simple quantum mechanical model, commonly known as the supersymmetric oscillator [16]. It consists of a bosonic and a fermionic oscillator of the same frequency. Therefore, we can write the Hamiltonian, for the system as

$$H = H_B + H_F = \omega \left( a_B^\dagger a_B + a_F^\dagger a_F \right) \tag{143}$$

where  $a_B$  and  $a_F$  describe, respectively, the bosonic and the fermionic annihilation operators.

The immediate thing to note from the structure of the Hamiltonian in eq. (143) is that there is no zero point energy. We will see this shortly as a general feature of supersymmetric theories. Let us also define two fermionic operators of the form

$$Q = a_B^\dagger a_F, \quad \bar{Q} = a_F^\dagger a_B \tag{144}$$

With the usual canonical commutation relations for the bosonic operators (see eq. (66)) and anti-commutation relations for the fermionic operators (see eq. (47)), it is easy to check that

$$[Q, H] = 0 = [\bar{Q}, H]$$

Namely, these fermionic operators are conserved. In fact, together with the Hamiltonian, they satisfy the algebra (it is straightforward to check this)

$$[Q, H] = 0 = [\bar{Q}, H]$$

$$\begin{aligned}
[Q, Q]_+ &= 0 = [\bar{Q}, \bar{Q}]_+ \\
[Q, \bar{Q}]_+ &= \frac{1}{\omega} H
\end{aligned}
\tag{145}$$

Such an algebra, where both commutators and anti-commutators are involved (or alternately, where there is a grading of the multiplication rule of the algebra), is known as a graded Lie algebra and supersymmetric theories are realizations of graded Lie algebras.

As we know from the study of symmetries, conserved quantities generate infinitesimal symmetries of the theory. Since both  $Q$  and  $\bar{Q}$  are conserved, it is worth asking what kind of symmetry transformations of the theory they generate. In fact, let us keep in mind that they are fermionic operators and hence the symmetry they will generate cannot be conventional. Explicitly, we can check that

$$\begin{aligned}
[Q, a_B^\dagger] &= 0 = [Q, a_F]_+ \\
[Q, a_B] &= -a_F \\
[Q, a_F^\dagger]_+ &= a_B^\dagger \\
[\bar{Q}, a_B] &= 0 = [\bar{Q}, a_F^\dagger]_+ \\
[\bar{Q}, a_B^\dagger] &= a_F^\dagger \\
[\bar{Q}, a_F]_+ &= a_B
\end{aligned}
\tag{146}$$

Namely,  $Q$  and  $\bar{Q}$  take bosonic operators to fermionic ones and *vice versa* which is the benchmark of supersymmetry. Thus, our Hamiltonian in eq. (143) is invariant under supersymmetric transformations of the form (146).

There are several things to note from the structure of the supersymmetry algebra in eq. (145). First, the energy eigenvalues of our supersymmetric theory have to be positive semi-definite since the operator on the left hand side of the last relation in (145) is. Furthermore, if the ground state is supersymmetric satisfying

$$Q|0\rangle = 0 = \bar{Q}|0\rangle \tag{147}$$

then, the ground state will have vanishing energy, as we had pointed out earlier as the case for our system. Both these results are, in fact, quite general for any supersymmetric theory. We also note from the structure of the algebra that the spectrum of the Hamiltonian will be doubly degenerate except for the ground state. Namely, if  $|\psi\rangle$  is an eigenstate of the Hamiltonian, then,  $Q|\psi\rangle$  (or,  $\bar{Q}|\psi\rangle$ ) – only one of them would be nontrivial depending on the form of  $|\psi\rangle$  – would be degenerate in energy.

Let us, in fact, examine some of these general results explicitly. The spectrum of the Hamiltonian in eq. (143) is, in fact, quite straightforward. The Hilbert space is a product space containing bosonic and fermionic oscillator states and a general state has the structure

$$|n_B, n_F\rangle = |n_B\rangle \otimes |n_F\rangle = \frac{(a_B^\dagger)^{n_B} (a_F^\dagger)^{n_F}}{\sqrt{n_B!}} |0, 0\rangle \tag{148}$$

with energy eigenvalues

$$E_{n_B, n_F} = \omega(n_B + n_F), \quad n_F = 0, 1; \quad n_B = 0, 1, 2, \dots \tag{149}$$

where the ground state is expected to satisfy

$$a_B|0\rangle = 0 = a_F|0\rangle \tag{150}$$

We note that an immediate consequence of (150) is that

$$Q|0\rangle = 0 = \bar{Q}|0\rangle$$

and, consequently, the ground state is supersymmetric and that the ground state energy is seen from (149) to vanish. All the higher states have positive energy. Furthermore, we note that all the states (except the ground state) of the form  $|n_B, 0\rangle$  and  $|n_B - 1, 1\rangle$  are degenerate in energy. Let us also note the effect of  $Q$  and  $\bar{Q}$  acting on the states of the Hilbert space, namely,

$$\begin{aligned} Q|n_B, n_F\rangle &= \begin{cases} \sqrt{n_B + 1}|n_B + 1, n_F - 1\rangle & \text{if } n_F \neq 0 \\ 0 & \text{if } n_F = 0 \end{cases} \\ \bar{Q}|n_B, n_F\rangle &= \begin{cases} \frac{1}{\sqrt{n_B}}|n_B - 1, n_F + 1\rangle & \text{if } n_B \neq 0 \text{ or } n_F \neq 1 \\ 0 & \text{if } n_B = 0 \text{ or } n_F = 1 \end{cases} \end{aligned} \quad (151)$$

### Supersymmetric Oscillator at $T \neq 0$

Let us next analyze the supersymmetric oscillator at finite temperature in the formalism of thermo field dynamics. As we had noted earlier, this is the ideal setting to discuss questions such as symmetry breaking. Let us note, even before carrying out the calculations, that we expect supersymmetry to be broken at finite temperature. Intuitively, this is quite clear. Namely, supersymmetry takes bosons to fermions and *vice versa* and, consequently, any boundary condition that distinguishes between the two would lead to a breaking of this symmetry. Temperature, in fact, introduces such a condition, namely, bosons and fermions behave differently at finite temperature (they obey distinctly different statistics). However, what is not clear *a priori* is whether such a breaking would be explicit or spontaneous.

To study the system at finite temperature within the framework of thermo field dynamics, let us look at the complete system, including the tilde oscillators, described by

$$\hat{H} = H - \tilde{H} = \omega(a_B^\dagger a_B + a_F^\dagger a_F) - \omega(\tilde{a}_B^\dagger \tilde{a}_B + \tilde{a}_F^\dagger \tilde{a}_F) \quad (152)$$

The Hilbert space of the doubled system has the structure

$$|n_B, n_F; \tilde{n}_B, \tilde{n}_F\rangle = |n_B, n_F\rangle \otimes |\tilde{n}_B, \tilde{n}_F\rangle \quad (153)$$

The thermal vacuum can now be defined (as discussed in section 3). Let us define

$$G(\theta_B, \theta_F) = -i\theta_B(\beta)(\tilde{a}_B a_B - a_B^\dagger \tilde{a}_B^\dagger) - i\theta_F(\beta)(\tilde{a}_F a_F - a_F^\dagger \tilde{a}_F^\dagger)$$

with (see eqs. (58) and (72))

$$\tan \theta_F(\beta) = e^{-\beta\omega/2} = \tanh \theta_B(\beta) \quad (154)$$

Then, the thermal vacuum can be defined as

$$|0, \beta\rangle = e^{-iG(\theta_B, \theta_F)} |0\rangle \quad (155)$$

This also allows us to calculate the thermal operators in a straightforward manner.

Let us note next that the expectation value of the Hamiltonian in the thermal vacuum is given by

$$\begin{aligned} E_0(\beta) &= \langle 0, \beta | H | 0, \beta \rangle = \langle 0, \beta | \omega(a_B^\dagger a_B + a_F^\dagger a_F) | 0, \beta \rangle \\ &= \omega(\sinh^2 \theta_B(\beta) + \sin^2 \theta_F(\beta)) = \frac{2\omega e^{-\beta\omega}}{(1 - e^{-2\beta\omega})} \end{aligned} \quad (156)$$

This shows that the energy of the thermal vacuum is nonzero for any finite temperature signaling that supersymmetry is broken. Furthermore, let us note that

$$\begin{aligned} Q|0, \beta\rangle &= a_B^\dagger a_F |0, \beta\rangle = \frac{e^{-\beta\omega/2}}{\sqrt{1 - e^{-2\beta\omega}}} |n_B(\beta) = 1, n_F(\beta) = 0; \\ \tilde{n}_B(\beta) &= 0, \tilde{n}_F(\beta) = 1\rangle \\ \bar{Q}|0, \beta\rangle &= a_F^\dagger a_B |0, \beta\rangle = \frac{e^{-\beta\omega/2}}{\sqrt{1 - e^{-2\beta\omega}}} |n_B(\beta) = 0, n_F(\beta) = 1; \\ \tilde{n}_B(\beta) &= 1, \tilde{n}_F(\beta) = 0\rangle \end{aligned} \quad (157)$$

This, in fact, shows that supersymmetry breaking is spontaneous at finite temperature and the new states on the right hand side of (157) would correspond to the appropriate quasi particle Goldstino states associated with such a symmetry breaking.

There are various other order parameters for the breaking of supersymmetry and all of them lead to the same conclusion that supersymmetry is spontaneously broken at finite temperature [16].

## 9 Conclusion

In this article, we have tried to describe some of the interesting features of finite temperature field theories. There are, of course, many more topics that we have not been able to discuss. However, it is our hope that the topics discussed, in this article, would raise the curiosity of the readers to pursue various other questions in this field.

This work was supported in part by the U.S. Dept. of Energy Grant DE-FG 02-91ER40685.

## References

- [1] T. Matsubara, *Prog. Theor. Phys.* **14** (1955) 351.
- [2] J. Schwinger, *J. Math. Phys.* **2** (1961) 407; J. Schwinger, *Lecture Notes Of Brandeis University Summer Institute* (1960).
- [3] H. Umezawa, H. Matsumoto and M. Tachiki, *Thero Field Dynamics and Condensed States*, North-Holland, Amsterdam, 1982.
- [4] D. A. Kirzhnits and A. D. Linde, *Phys. Lett.* **42B** (1979) 471; L. Dolan and R. Jackiw, *Phys. Rev.* **D9** (1974) 3320; S. Weinberg, *Phys. Rev.* **D9** (1974) 3357.
- [5] D. J. Gross, R. D. Pisarski and L. G. Yaffe, *Rev. Mod. Phys.* **53** (1981) 43.
- [6] A. A. Anselm, *Phys. Lett.* **B217** (1989) 169; A. A. Anselm and M. G. Ryskin, *Phys. Lett.* **B266** (1991) 482; J. D. Bjorken, *Int. J. Mod. Phys.* **A7** (1992) 4189; J. P. Blaizot and A. Krzywicki, *Phys. Rev.* **D46** (1992) 246; K. Rajagopal and F. Wilczek, *Nuc. Phys.* **B204** (1993) 577; P. F. Bedaque and A. Das, *Mod. Phys. Lett.* **A8** (1993) 3151.
- [7] A. Das, *Field Theory, A Path Integral Approach*, World Scientific (1993).
- [8] A. L. Fetter and J. D. Walecka, *Quantum Theory of Many Particle Systems*, McGraw-Hill (1971); A. A. Abrikosov, L. P. Gorkov and I. E. Dzyaloshinski, *Methods of Quantum Field Theory in Statistical Physics*, Dover (1975).
- [9] J. I. Kapusta, *Finite Temperature Field Theory*, Cambridge University Press (1989); M. Le Bellac, *Thermal Field Theory*, Cambridge University Press (1996).
- [10] A. Das, *Finite Temperature Field Theory*, World Scientific (1997).
- [11] N. P. Landsman and C. G. van Weert, *Phys. Rep.* **145** (1987) 141.
- [12] P. M. Bakshi and K. T. Mahanthappa, *J. Math. Phys.* **4** (1963) 1; L. V. Keldysh, *Sov. Phys. JETP* **20** (1965) 1018; K. C. Chou et al, *Phys. Rep.* **118** (1985) 1.
- [13] H. A. Weldon, *Phys. Rev.* **D47** (1993) 594; P. F. Bedaque and A. Das, *Phys. Rev.* **D47** (1993) 601.
- [14] G. Dunne, K. Lee and C. Lu, *Phys. Rev. Lett.* **78** (1997) 3434.
- [15] A. Das and G. Dunne, *Phys. Rev.* **D57** (1998) 5023.

- [16] A. Das and M. Kaku, *Phys. Rev.* **D18** (1978) 4540; A. Das, A. Kharev and V. S. Mathur, *Phys. Lett.* **B181** (1986) 299; A. Das and V. S. Mathur, *Phys. Rev.* **D35** (1987) 2053; A. Das, *Physica* **A158** (1989) 1.
- [17] R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals*, McGraw-Hill (1965).
- [18] L. S. Shulman, *Techniques and Applications of Path Integration*, John-Wiley (1981); H. Kleinert, *Path Integrals*, World Scientific (1995).
- [19] F. A. Berezin, *The Method of Second Quantization*, Academic Press (1966); B. DeWitt, *Supermanifolds*, Cambridge University Press (1984).
- [20] R. Kubo, *J. Phys. Soc. Japan*, **12** (1957) 570; P. Martin and J. Schwinger, *Phys. Rev.* **115** (1959) 1342.
- [21] C. Bloch, *Nuc. Phys.* **7** (1958) 451.
- [22] A. A. Abrikosov, L. P. Gorkov and I. E. Dzyaloshinski, *Sov. Phys. JETP* **9** (1959) 636; H. Umezawa, Y. Tomozawa and H. Ezawa, *Nuovo Cim.* **5** (1957) 810.
- [23] H. A. Weldon, *Phys. Rev.* **D26** (1982) 1394; *ibid* **D28** (1983) 2007.
- [24] G. Baym and N. Mermin, *J. Math. Phys.* **2** (1961) 232.
- [25] Y. Takahashi and H. Umezawa, *Collective Phenomena* **2** (1975) 55.
- [26] I. Ojima, *Ann. Phys.* **137** (1981) 1.
- [27] A. J. Niemi and G. Semenoff, *Ann. Phys.* **152** (1984) 105.
- [28] P. F. Bedaque and A. Das, *Phys. Rev.* **D45** (1992) 2906.
- [29] A. Das and M. Hott, *Mod. Phys. Lett.* **A9** (1994) 3383.
- [30] S. Deser, R. Jackiw and S. Templeton, *Ann. Phys.* **140** (1982) 372.
- [31] K. S. Babu, A. Das and P. Panigrahi, *Phys. Rev.* **D36** (1987) 3725; I. Aitchison and J. Zuk, *Ann. Phys.* **242** (1995) 77; N. Bralić, C. Fosco and F. Schaposnik, *Phys. Lett.* **B383** (1996) 199; D. cabra, E. Fradkin, G. Rossini and F. Schaposnik, *Phys. Lett.* **B383** (1996) 434.
- [32] G. Dunne, R. Jackiw and C. Trugenberger, *Phys. Rev.* **D41** (1990) 661.
- [33] S. Deser, L. Griguolo and D. Seminara, *Phys. Rev. Lett.* **79** (1997) 1976; C. Fosco, G. Rossini and F. Schaposnik, *Phys. Rev.* **79** (1997) 1980; S. Deser, L. Griguolo and D. Seminara, *Phys. Rev.* **D57** (1998) 7444; C. Fosco, G. Rossini and F. Schaposnik, *Phys. Rev.* **D56** (1997) 6547.
- [34] Y. A. Gel'fand and E. P. Likhtman, *JETP Lett.* **13** (1971) 323; P. Ramond, *Phys. Rev.* **D3** (1971) 2415; A. Neveu and J. Schwarz, *Nuc. Phys.* **31** (1971) 86; D. Volkov and V. Akulov, *Phys. Lett.* **B46** (1974) 109; J. Wess and B. Zumino, *Nuc. Phys.* **70** (1974) 39.
- [35] P. Fayet and S. Ferrara, *Phys. Rep.* **32** (1977) 249; M. F. Sohnius, *Phys. Rep.* **128** (1985) 39.

# 16. Integrable Models And The Toda Lattice Hierarchy

Bani M Sodermark \*

Dept of Engineering Sciences, Physics and Mathematics,  
Karlstad University, 65188 Karlstad, SWEDEN

## Abstract

A pedagogical presentation of integrable models with special reference to the Toda lattice hierarchy has been attempted. The example of the *KdV* equation has been studied in detail, beginning with the infinite conserved quantities and going on to the Lax formalism for the same. We then go on to symplectic manifolds for which we construct the Lax operator. This formalism is applied to Toda Lattice systems. The Zakharov Shabat formalism aimed at encompassing all integrable models is also covered after which the zero curvature condition and its fallout are discussed. We then take up Toda Field Theories and their connection to W algebras via the Hamiltonian reduction of the WZNW model. Finally, we dwell on the connection between four dimensional Yang Mills theories and the *KdV* equation along with a generalization to supersymmetry.

## 1 Introduction: Non-Linear Equations

Linear partial differential equations, in particular the Schroedinger, Klein-Gordon and Dirac equations, have been known in field theory over a long time, and have been used in many different problems with great success. Non-linear equations, i.e., equations where the potential term is non-linear in the field ( $S$ ), have been known for some time as well. These equations and their solutions are the topic of the present Article.

The earliest non-linear wave equations known in physics were the Liouville and Sine-Gordon equations. The Liouville equation arose in the context of a search for a manifold with constant curvature. Pictorially, such parametrizations may be likened to covering a surface with a fishing net. Since the knots on the fishing net do not move, the arc length is constant. The threads in the net correspond to a local coordinate system on the surface.

The Liouville manifolds may be reparametrized locally so as to have a metric of the form:

$$A = \begin{pmatrix} \exp \rho & 0 \\ 0 & \exp \rho \end{pmatrix} \quad (1.1)$$

so as to be conformally equivalent to a flat space metric. The study of such manifolds with constant curvature led J.Liouville [1] to the equation known by his name:

$$\frac{\partial^2 \rho}{\partial x \partial y} = \exp \rho \quad (1.2)$$

$x$  and  $y$  being local orthogonal coordinates. Interest in this equation was renewed in the 70's and 80's due to its appearance in string theories [2,3,4].

The Sine-Gordon equation, named after a pun on the Klein-Gordon equation, is an equation for the angle  $\omega$  between two coordinate lines when the total curvature is constant and negative.

---

\*Email: bani.sodermark@kau.se

This equation first appeared in the work of Enneper in 1870, and has the form:

$$\frac{\partial^2 \omega}{\partial x \partial y} = \sin(\omega) \quad (1.3)$$

where  $x$  and  $y$  are coordinates in a system with constant arc length.

The Sine-Gordon equation has some interesting solutions known as *solitons* and *breathers*. A soliton satisfies three conditions. First, a single soliton must have constant shape and velocity. Secondly, it must be localized, and its derivative must vanish at infinity. Thirdly, if two solutions collide, they should survive the collision with their shapes unchanged.

Principally, there are two types of solitons, one which increases by a fixed amount (say  $2\pi$ ), and is called a 'kink'; the other which decreases by the same amount, and is called an 'anti-kink'.

A breather is a localized solution that varies periodically, and could be considered as a permanently bound system of a kink and anti-kink.

An interesting property of the Sine-Gordon equation is that its solutions can be mapped into others through the Baecklund transformation [5], and can thus be used to create new solutions from known solutions. It is however impossible to generate a complete set of solutions from one original solution, via the Baecklund transformation [5].

A third non-linear equation which we shall study in some detail, was discovered in 1895 by D.J.Korteweg and G. de Vries [6], while trying to describe the motion of water-waves in a canal. It has the form:

$$u_t - 6uu_x + u_{xxx} = 0 \quad (1.4)$$

and is also known as the *KdV* equation. It has been extensively studied, and many of the properties of non-linear wave equations that are known today, were discovered in connection with its solution. This equation was solved by Gardner, Greene, Kruskal, and Miura in 1967 [7-13]. Along with N.J. Zabusky and C. H. Su, they also found many interesting properties of the same. One of these is that the *KdV* equation has an infinite number of conservation laws, and that the conserved quantities of each of these laws can be used as a Hamiltonian for an integrable system. This collection of Hamiltonians is called the *KdV* hierarchy.

There exists a theorem of classical mechanics, which states that if a Hamiltonian system with  $2n$  degrees of freedom has  $n$  functionally independent conserved quantities such that the Poisson bracket of any two of them vanishes, i.e., the integrals of motion are in 'involution', the system is completely integrable. It is clear that solutions of systems with an infinite number of conserved quantities must be infinitely restricted. A soliton is precisely such a solution: it is a localized wave which retains its shape even after collisions. Intuitively, it is clear that for this to happen, there must be an infinite number of conservation laws, and therefore an infinite number of conserved quantities. The terms 'integrable models' and 'solitons' are often used synonymously.

A system of coupled equations of motion describing a 1-dimensional crystal with non-linear coupling between nearest neighbour atoms, was introduced by M.Toda [14] in 1967. The equations of motion are

$$m \frac{d^2 r_n}{dt^2} = a[2e^{-r_n} - e^{-r_{n-1}} - e^{-r_{n+1}}] \quad (1.5)$$

where  $r_n = u_{n+1} - u_n$ , and  $u_n(t)$  is the longitudinal displacement of the  $n$ -th atom with mass  $m$  from its equilibrium position,  $a$  being a constant. These models admit soliton solutions which have been studied experimentally on an electrical network by K. Hitota and K. Suzuki [15]. In the continuum limit, these equations reduce to the *KdV* equation [5].

We see that models with exponential interactions are a source of non-linear equations, the Liouville and Sine-Gordon equations being examples. The Liouville equation could be generalized to include a mass term:

$$\frac{\partial^2 \phi}{\partial x \partial y} + m^2 \phi = e^\phi \quad (1.6)$$

while the Sine-Gordon equation could be generalized to the "Sinh-Gordon" equation with the replacement  $\omega \rightarrow i\omega$ . Thus

$$\frac{\partial^2 \omega}{\partial x \partial y} + m^2 \omega = \sinh \omega \quad (1.7)$$



We also have the Toda Field Theory equations

$$\frac{\partial^2 \phi_i}{\partial x \partial y} = -e^{k_{ij} \phi_j} \quad (1.8)$$

Here  $k_{ij}$  is the Cartan matrix for some complex Lie Algebra. The simplest of these field theories is the  $A_r$  Toda field theory, and it includes the Liouville field theory for the special case  $r = 1$ . There exist generalizations of the Toda equations called "Affine Toda Equations", and have an extra term on the RHS, taking the form:

$$\frac{\partial^2 \phi_i}{\partial x \partial y} = -e^{k_{ij} \phi_j} + \gamma R_i e^{k_0 \phi_j} \quad (1.9)$$

Here  $K$  is an affine Cartan matrix, and  $R_i$  the right null vector for this matrix when  $R_0$  is normalized to unity.

These models include the Sinh-Gordon equation as a special case. Both the Toda and Affine Toda field theories have an infinite number of conserved quantities [16]. They admit soliton solutions with an imaginary  $\phi_i$  [17]. Both models have been formally solved by Leznov and Saveliev [18].

The Toda field theories can be obtained from the Toda Lattice by setting

$$\psi_i = (\phi_i - \phi_{i-1}) - (\phi_{i+1} - \phi_i) \quad (1.10)$$

whence

$$\frac{\partial^2 \psi_i}{\partial t^2} - \frac{\partial^2 \psi_i}{\partial x^2} = -[2e^{\psi_i} - e^{\psi_{i-1}} - e^{\psi_{i+1}}] \quad (1.11)$$

for  $SU(n+1)$ , showing that the space-independent solutions of (1.11) satisfy (1.5).

Since the Toda field theories are the  $\gamma = 0$  limits of the Affine Toda field theories, they could be used to classify 2-dimensional models with a second order phase transition, with the Toda field theory describing the model at the critical point where it has to be conformally invariant [19]. Hence the great interest in (Affine) Toda field theories. However the precise connection is still unclear. Central charges and critical exponents have been calculated and compared. One hopes that the Affine Toda field theories are perturbations that correspond to the physical model away from the critical point. However, more explicit connections are yet to be found.

The method originally used for solving non-linear equations, and especially the  $KdV$  equation, was the inverse scattering method originated by Gelfand and Levitan [20]. This involved looking for a linear equation related to the original non-linear equation, and studying the evolution of the latter. In 1968, P.Lax provided this method within a solid theoretical framework [21]. The Lax equation is

$$L_t + [L, M] = 0 \quad (1.12)$$

where  $L$  and  $M$  are operators satisfying

$$L\psi = \lambda\psi; \quad (1.13)$$

and

$$\psi_t = M\psi \quad (1.14)$$

where  $\lambda$  is a scalar, and  $\psi$  a solution of a linear equation which is just the Schroedinger equation for the  $KdV$  case ! The Lax equation was generalized to the form of a zero curvature condition which facilitates greatly the form of the transition matrix from the initial to the final state.

In what follows, we attempt to give a pedagogical presentation of Integrable Systems with special emphasis on the  $KdV$  and Toda systems. After an introduction to the  $KdV$  equation and its properties, we show how an infinite number of conserved quantities arise via the Muira [8] transformation, while detailed calculations are referred to ref.[22]. We then dwell on solutions of the  $KdV$  equation via the inverse scattering method and the Lax formalism [21], after which we obtain the Lax operator for symplectic manifolds, using the Toda Lattice as an example. The

group structure of the Toda equations for  $SU(N)$  is also studied. The Lax transformation was later generalized by Zakharov and Shabat [23] to a first order formalism which was used by Ablowitz, Kamp, Newell and Segur (AKNS) [24], for a unified description of other integrable models. The essential features of this approach are also discussed. A fall-out of the above is the ‘zero curvature condition’ that facilitates the transition to the quantum case. However, the treatment we follow is strictly classical.

Next we take up the Toda field theories, and after reporting briefly the connection with conformal invariance, dwell on the Hamiltonian reduction of the WZNW model to the Toda field theory, which in effect transforms an affine Lie Algebra to a W-Algebra. (Most calculational details are skipped, but may be found in the literature [25]). Finally we refer to the interesting connection between the 4D self-dual Yang-Mills theory and 2D Integrable models, and the generalization to SuperSymmetry.

The material is presented as follows. In Sect.2, we introduce the  $KdV$  equation and its conserved quantities. In Sect.3, solutions of non-linear equations are taken up, in particular the inverse scattering method and the Lax formalism. In Sect.4, we digress to Symplectic Manifolds and construct conserved quantities for these manifolds. Sect.5 applies the above framework to the Toda Lattice where the group structure of the Toda equations is also discussed. In Sect.6, we take up the unifying first order formalism of Zakharov and Shabat [23], continuing in Sect.7 to the zero curvature formalism and its ramifications. In Sect.8, we take up Conformal Invariance, and introduce Toda Field Theories which are constructed independently of the Toda Lattice. In Sect.9, we carry out the Hamiltonian reduction of the WZNW model to Toda Field Theories. Finally in Sect.10, we take up the connection of Toda Field Theories with Self-dual Yang-Mills models. Sect.11 contains some concluding remarks.

## 2 The $KdV$ Equation

The  $KdV$  equation was formulated to explain the solitary water waves observed by J.Scott Russell in the Edinburgh Glassgow canal. It is a non-linear equation in one space and one time dimension and possesses soliton solutions. Of this, however, nothing was known at the time of its formation.

The  $KdV$  equation after an initial scaling takes the form

$$\frac{\partial u}{\partial t} = u \frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3} \quad (2.1)$$

This equation is Galilean invariant, but not Lorentz invariant. It can be derived from the Hamiltonian

$$H(u) = \int_{-\infty}^{+\infty} \left[ \frac{u^3}{6} - \frac{1}{2} \left( \frac{\partial u}{\partial x} \right)^2 \right] dx \quad (2.2)$$

where the  $u(x)$  satisfy the Poisson bracket relations

$$[u(x), u(y)] = \partial_x \delta(x - y) \quad (2.3)$$

However, the Lagrangian from which it can be derived, is non local:

$$L_{KdV} = \frac{1}{2} \int_{-\infty}^{+\infty} dx dy u(x) \epsilon(x - y) \frac{\partial u(y)}{\partial t} - \int dx \left[ \frac{u^3}{6} - \frac{1}{2} \left( \frac{\partial u}{\partial x} \right)^2 \right] \quad (2.4)$$

where

$$\epsilon(x - y) = \theta(x - y) - \frac{1}{2}, \quad (2.5)$$

$\theta$  being the step function. Ergo, one cannot write down a local Lagrangian whose Euler-Lagrange equations yield the  $KdV$  equation.

Solutions of the  $KdV$  equation can be shown to be soliton solutions which travel without any change of shape. It is the non-linear term which is responsible for the above property.

What is most interesting about the  $KdV$  equation is that it admits of an infinite number of conserved quantities as was shown by Miura [8]. This procedure is explained below.

The  $KdV$  equation is related to another equation called the modified  $KdV$  ( $MKdV$ ) equation, viz.,

$$\frac{\partial v}{\partial t} = v^2 \frac{\partial v}{\partial x} + \frac{\partial^3 v}{\partial x^3} \quad (2.6)$$

where  $v$  is related to  $u$  in the  $KdV$  equation through the Riccati transformation

$$u = v^2 + i\sqrt{6} \frac{\partial v}{\partial x} \quad (2.7)$$

The  $MKdV$  equation is however not Galilean invariant. Under the transformation

$$t \rightarrow t; \quad x \rightarrow x + \frac{3t}{2\epsilon^2}; \quad u \rightarrow u + \frac{3}{2\epsilon^2}; \quad v \rightarrow \frac{v\epsilon}{\sqrt{6}} + \frac{\sqrt{6}}{2\epsilon} \quad (2.8)$$

it reduces to

$$\partial_t v = \left( \frac{\epsilon^2 v^2}{6} + v \right) \partial_x v + \partial_x^3 v = \partial_x \left[ \frac{\epsilon^2 v^3}{18} + \frac{v^2}{2} + \partial_x^2 v \right] \quad (2.9)$$

This yields a solution of the  $KdV$  equation through the transformation

$$u = \epsilon^2 v^2 / 6 + v + i\epsilon \partial_x v \quad (2.10)$$

The second form of (2.9) is in the nature of a continuity equation, so that we can identify

$$K = \int_{-\infty}^{+\infty} dx v(x(t)) \quad (2.11)$$

as the conserved quantities.  $v$  can be inverted in terms of  $u$  as

$$v = \sum_0^{\infty} \epsilon^n v_n(u(x, t)) \quad (2.12)$$

and this yields  $v_n(u(x, t))$  as the conserved densities, since each power of  $\epsilon$  must independently satisfy a continuity equation. That these are also in involution can also be checked, being explicitly shown by Das [22]. Some of the conserved quantities are

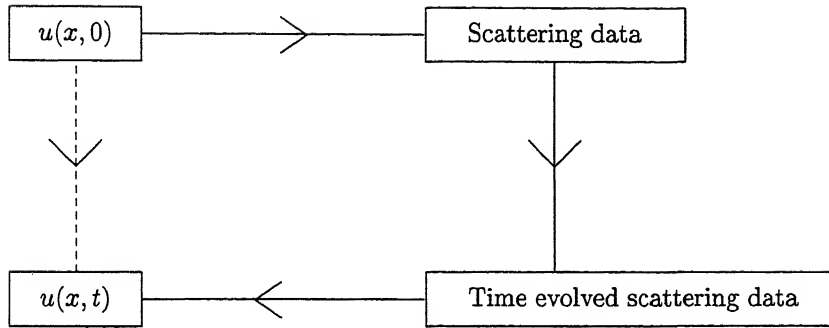
$$v_1 = -i\partial_x u_1; \quad v_2 = -\frac{u^2}{6} - \partial_x^2 u; \quad v_3 = i\partial_x \left[ \frac{u^2}{3} + \partial_x^2 u \right] \quad (2.13)$$

### 3 The Lax Framework

Linear Hamiltonian systems with fixed initial value problems can be solved using the Laplace or Fourier transformations. Such methods are inapplicable for the nonlinear equations and new methods must be found. Gardner, Green, Krushal and Miura [9] managed to solve the initial value problem for the  $KdV$  equation in a very ingenious way. In subsequent years, this method has become the standard method for solving non-linear systems and goes by the name of inverse scattering theory [20,21]. This method is outlined in Fig 1.

The initial value for the partial differential equation is used as the potential in a 1-dimensional scattering problem for a linear equation, e.g. the Schroedinger equation. One then finds the so called scattering data, i.e. discrete spectrum, normalization constants, reflection constants (as a function of the wave number) for this scattering problem. Using the partial differential equation ( $pde$ ) evaluated for  $|x|$  asymptotically large, (and hence the  $pde$  becomes a linear equation because the potential is assumed to vanish at spatial infinity), the values of the scattering data can be found for all later times. Finally, the scattering data allow one to reconstruct the potential, and hence the solution of the  $pde$  for any later time.

Figure 1:



One would intuitively like a better understanding of the origin and relevance of the linear Schroedinger equation. One way to see this is through a generalized Riccati relation of the form:

$$u + 6\lambda = v^2 + i\sqrt{6}\frac{\partial v}{\partial x} \quad (3.1)$$

so that the  $KdV$  relation (2.1) reduces to

$$\frac{\partial v}{\partial t} - (v^2 - 6\lambda)\frac{\partial v}{\partial x} - \frac{\partial^3 v}{\partial x^3} = 0 \quad (3.2)$$

As mentioned earlier, a solution of the  $MKdV$  equation yields a solution of the  $KdV$  equation through the Riccati relation. The simplest way to attempt an inversion of the Riccati relation is to linearize it. To that end we define

$$v = i\sqrt{6}\psi_x/\psi \quad (3.3)$$

so that (3.1) takes the form

$$u + 6\lambda = -6\psi_{xx}/\psi, \quad (3.4)$$

or equivalently,

$$\psi_{xx} + \left(\frac{u}{6} + \lambda\right)\psi = 0 \quad (3.5)$$

which is, in fact, the time-independent Schroedinger equation. There exists however a more formal theory due to Lax [21], which we now elaborate.

Given a linear equation described by a time-independent Hamiltonian  $H$ , and an operator  $A$  whose expectation values are time independent,  $A(t)$  is unitarily equivalent to  $A(0)$ :

$$U^\dagger(t)A(t)U(t) = A(0) \quad (3.6)$$

where  $U(t)$  is the time-evolution operator with the form

$$U(t) = \exp[-iHt] \quad (3.7)$$

Differentiating (3.6) gives

$$U^\dagger(t)\left(\frac{\partial A}{\partial t} - i[A, H]\right)U(t) = 0 \quad (3.8)$$

which implies

$$\frac{\partial A}{\partial t} = i[A, H] \quad (3.9)$$

Thus for the expectation value of  $A(t)$  to be time independent, the standard time evolution relation (3.8) must be satisfied. Further, from eq.(3.7) follows the relation

$$\frac{\partial U(t)}{\partial t} = -iHU(t) = BU(t) \quad (3.10)$$

where

$$B = -iH \quad (3.11)$$

is an anti-Hermitian operator.

This argument is mimicked in the case of a non-linear evolution equation. Let

$$L(u(x, t)) = L(t) \quad (3.12)$$

denote the linear operator we seek. We assume it to be Hermitian, and to have eigen-values independent of  $t$ . For this to be true, one must have  $u^\dagger(t)L(t)u(t)=L(0)$ . Differentiating both sides w.r.t.  $t$ , we obtain

$$\frac{\partial U^\dagger(t)}{\partial t}L(t) + U^\dagger(t)\frac{\partial L(t)}{\partial t}U(t) + U^\dagger(t)L(t)\frac{\partial U(t)}{\partial t} \quad (3.13)$$

Unlike the linear case, we do not know the form of  $U(t)$ . However,  $U$  is unitary, so

$$U^\dagger U = 1 \Rightarrow \frac{\partial U^\dagger(t)}{\partial t}U(t) + U^\dagger\frac{\partial U(t)}{\partial t} = 0 \quad (3.14)$$

Thus we can write

$$\frac{\partial U(t)}{\partial t} = B(t)U(t) \quad (3.15)$$

where anti-hermiticity must be imposed on  $B$ . Substitution in (3.12), and a little simplification, yields

$$\frac{\partial L(t)}{\partial t} = [B(t), L(t)] \quad (3.16)$$

which is similar to (3.8), except for the fact that we do not yet know the form of  $B$ . However, let us assume that  $L(t)$  is linear in  $u(x, t)$ . Consequently, the LHS of (3.14) is a multiplicative operator, proportional to the time evolution operator of  $u(x, t)$ . This would ensure that the eigen-values  $\lambda$  of  $L(t)$  would be time-independent, i.e.,

$$L(t)\psi(t) = -\lambda\psi(t) \quad (3.17)$$

Further,  $\psi(t)$  must be unitarily related to its value at  $t = 0$ , i.e.,

$$\psi(t) = U(t)\psi(0), \quad (3.18)$$

and its evolution w.r.t. time would take the form

$$\frac{\partial \psi(t)}{\partial t} = \frac{\partial U(t)}{\partial t}\psi(0) = B(t)\psi(t) \quad (3.19)$$

The operators  $L(t)$  and  $B(t)$ , when they exist, are known as the Lax pair, corresponding to a given non-linear evolution equation, and play a fundamental role in determining the solution. For the KdV equation,  $L(t)$  is obtained from the linear form of the Schroedinger equation

$$L(t) = D^2 + \frac{1}{6}u(x, t); \quad D \equiv \frac{\partial}{\partial x}. \quad (3.20)$$

By trial and error,  $B(t)$  can be chosen so that (3.15) is satisfied, and a possible solution is

$$B(t) = 4D^3 + \frac{1}{2}(Du + uD) \quad (3.21)$$

The solution for  $\psi$  w.r.t.  $t$  follows from (3.18) and (3.20) to be

$$\psi_t = 4\psi_{xxx} + \frac{1}{2}u_x\psi + u\psi_x + \text{const.}\psi \quad (3.22)$$

which yields, using the Schroedinger equation (3.5):

$$\psi_t + \frac{1}{6}u_x\psi - \frac{1}{3}u\psi_x + 4\lambda\psi_x = \text{const.}\psi \quad (3.23)$$

A. Lenard [26], in an unpublished report, further displayed the relation between the Schroedinger equation and the  $KdV$  relation by elegantly deriving the latter from the former, using only the assumption that the spectral parameter  $\lambda$  in (3.4) is time-independent.

The  $KdV$  equation exhibits also a fascinating symmetry, i.e., that of the group  $SL(2, R)$ . Consider a group element

$$g = \exp[i\theta^a T_a] \quad (3.24)$$

where  $T_a$  is a generator of  $SL(2, R)$ , and define

$$A_\mu \equiv g^{-1} \partial_\mu g \quad (3.25)$$

Then the  $KdV$  equation follows from the fact that the Maurer-Cartan equation

$$\partial_\mu A_\nu - \partial_\nu A_\mu - [A_\mu, A_\nu] = 0 \quad (3.26)$$

is satisfied for a special for a special choice of gauge, e.g.,

$$A_1^1 = -\sqrt{\lambda}; (\lambda < 0); \quad A_1^3 = 6; \quad A_1^2 = -\frac{1}{36}u(x, t); \quad A_0^3 = A(u(x, t)) \quad (3.27)$$

## 4 Lax Formalism On Symplectic Manifolds

In this Section, we conclude the above study of the  $KdV$  equation with a with a short discussion on symplectic geometry, which is directly relevant for application to the Toda Lattice.

A symplectic manifold is one with a preferred 2-form  $f_{\mu\nu}$  which is non-degenerate and closed. The phase space of an integrable model corresponds to a very special symplectic manifold, since it possesses a dual Poisson bracket structure. We assume that there exist two distinct 2-forms which are both non-degenerate and closed. One way of expressing the existence of two distinct symplectic structures is to require that the same dynamical equation be described by two distinct first order Lagrangians  $L_0$  and  $L$ , where

$$L_0 = \theta_\mu^{(0)}(y) \dot{y}^\mu - H_0(y); \quad (4.1)$$

$$L = \theta_\mu(y) \dot{y}^\mu - H(y) \quad (4.2)$$

where

$$\dot{y}^\mu = \frac{dy^\mu}{dt}; \quad [\mu = 1, 2, \dots, 2N] \quad (4.3)$$

The Euler-Lagrangian equations following from (4.1-2) are

$$f_{\mu\nu}(y) \dot{y}^\nu = \partial_\mu H_0(y) \quad (4.4)$$

$$F_{\mu\nu}(y) \dot{y}^\nu = \partial_\mu H(y) \quad (4.5)$$

where

$$f_{\mu\nu} = \partial_\mu \theta_\nu^{(0)}(y) - \partial_\nu \theta_\mu^{(0)}(y) \quad (4.6)$$

$$F_{\mu\nu} = \partial_\mu \theta_\nu(y) - \partial_\nu \theta_\mu(y) \quad (4.7)$$

It is easy to see that the two forms  $f$  and  $F$  are closed, where

$$f = \frac{1}{2} f_{\mu\nu} dy^\mu \wedge dy^\nu \quad (4.8)$$

$$F = \frac{1}{2} F_{\mu\nu} dy^\mu \wedge dy^\nu \quad (4.9)$$

since  $f_{\mu\nu}$  and  $F_{\mu\nu}$  satisfy the Bianchi identities

$$\partial_\lambda f_{\mu\nu} + \partial_\mu f_{\nu\lambda} + \partial_\nu f_{\lambda\mu} = 0; \quad (4.10)$$

and

$$\partial_\lambda F_{\mu\nu} + \partial_\mu F_{\nu\lambda} + \partial_\nu F_{\lambda\mu} = 0. \quad (4.11)$$

Besides, they must also be non-degenerate since (4.4) and (4.5) describe the same dynamical system. Let their universes be  $f^{\mu\nu}$  and  $F^{\mu\nu}$ , i.e.,

$$f_{\mu\nu} f^{\nu\lambda} = F_{\mu\nu} F^{\nu\lambda} = \delta_\mu^\lambda \quad (4.12)$$

so that (4.4-5) take the forms

$$\dot{y}^\nu = f^{\nu\mu} \partial_\mu H_0(y) \quad (4.13)$$

$$\dot{y}^\nu = F^{\nu\mu} \partial_\mu H(y). \quad (4.14)$$

We can also construct a nontrivial  $(1, 1)$  tensor  $S_\mu^\nu$  as

$$S_\mu^\nu = F_{\mu\lambda}(y) f^{\lambda\nu}(y). \quad (4.15)$$

Consistency of (4.4) and (4.5) further requires that

$$\partial_\mu \partial_\nu H_0(y) - \partial_\nu \partial_\mu H_0(y) = 0 \quad (4.16)$$

so that after a little algebra, one can show that

$$\frac{df_{\mu\nu}(y)}{dt} = -U_\mu^\lambda f_{\lambda\nu} + U_\nu^\lambda f_{\lambda\mu} \quad (4.17)$$

where

$$U_\mu^\nu = \partial_\mu y^\nu = \partial_\mu [f^{\nu\lambda} \partial_\lambda H_0(y)] = \partial_\mu [F^{\nu\lambda} \partial_\lambda H(y)] \quad (4.18)$$

with a corresponding relation for  $F^{\mu\nu}$ , i.e.,

$$\frac{dF_{\mu\nu}(y)}{dt} = -U_\mu^\lambda F_{\lambda\nu} + U_\nu^\lambda F_{\lambda\mu} \quad (4.19)$$

involving the same  $U$ -tensor. The corresponding equations for the inverses  $f^{\mu\nu}$  and  $F^{\mu\nu}$  follow from (4.17) and (4.19), and have the forms

$$\frac{df^{\mu\nu}}{dt} = f^{\mu\lambda} U_\lambda^\nu - f^{\nu\lambda} U_\lambda^\mu \quad (4.20)$$

$$\frac{dF^{\mu\nu}}{dt} = F^{\mu\lambda} U_\lambda^\nu - F^{\nu\lambda} U_\lambda^\mu \quad (4.21)$$

We can finally show that

$$\frac{dS_\mu^\nu}{dt} = S_\mu^\lambda U_\lambda^\nu - U_\mu^\lambda S_\lambda^\nu \quad (4.22)$$

which in matrix notation

$$\frac{dS}{dt} = [S, U] \quad (4.23)$$

can be recognized as a Lax equation (3.15), thus providing a Lax representation of the dynamical equations (4.13) and (4.14). One important consequence of (4.23) is that the set of quantities

$$K_n = \frac{1}{n} \text{Tr} S^n \quad (4.24)$$

and

$$K_0 = \ln |\det S| \quad (4.25)$$

can be shown to be invariants since

$$\frac{dK_n}{dt} = \text{Tr} \left[ P(S) \frac{dS}{dt} \right] = \text{Tr} [P(S)[S, U]] = 0 \quad (4.26)$$

$P(S)$  is a polynomial in  $S$ . That these are in involution can easily be checked, as done explicitly in ref.[22]. Applied to the  $KdV$  equation, the two Poisson structures of that equation are given by the correspondence:

$$F^{\mu\nu} \rightarrow D; \quad (4.27)$$

$$f^{\mu\nu} \rightarrow D^3 + \frac{1}{3}(Du + uD) \quad (4.28)$$

Going to the coordinate bases we have

$$F(x, y) = \langle y | D | x \rangle = \partial_x \delta(x - y) \quad (4.29)$$

$$f(x, y) = \frac{\partial^3}{\partial x^3} + \frac{1}{3}(\partial_x u + u \partial_x) \delta(x - y) \quad (4.30)$$

so that

$$F^{-1}(x, y) = \epsilon(x - y) = \theta(x - y) - \frac{1}{2} \quad (4.31)$$

However  $f^{-1}(x - y)$  cannot be expressed in a closed form. The Lax operator  $S$  takes the form

$$S = D^2 + \frac{2}{3}u + \frac{1}{3}(Du)D^{-1}, \quad (4.32)$$

and with a little algebra, (4.23) can be shown to be reduced to the  $KdV$  equation, with consequently an infinite # of conserved quantities. This is described in detail in ref.[22].

## 5 The Toda Lattice

The model of the  $KdV$  equation that has been studied so far is a continuum model. A finite dimensional system with a finite # of degrees of freedom is simpler to study. The Toda Lattice is such a system to which the symplectic approach of the above Section is especially applicable. We now study the Toda Lattice and its integrability from a symplectic point of view, following it up with a group theoretical treatment.

The Toda Lattice describes the motion of  $N$  point masses on the line, under the influence of an exponential interaction. The Hamiltonian equations in terms of the canonical coordinates  $Q_i$  and momenta  $P_i$  are given by

$$\begin{aligned} \dot{Q}_i &= P_i; & (i = 1, 2, \dots, N); \\ \dot{P}_j &= e^{-(Q_j - Q_{j-1})} - e^{-(Q_{j+1} - Q_j)}; & (j = 2, 3, \dots, N-1); \\ \dot{P}_1 &= -e^{-(Q_2 - Q_1)}; & \dot{P}_N = e^{-(Q_N - Q_{N-1})}. \end{aligned} \quad (5.1)$$



The equations can be cast into a more symmetrical form by enlarging the system to  $(N + 2)$  point masses, with end points at spatial infinity. In that case, the Hamiltonian equations take the form :

$$\dot{Q}_i = P_i; \quad (i = 1, 2, \dots, N); \quad \dot{P}_i = e^{-(Q_i - Q_{i+1})} - e^{-(Q_{i+1} - Q_i)}. \quad (5.2)$$

We can choose

$$y^i = Q_i; \quad y^{N+i} = P_i; \quad (i = 1, 2, \dots, N). \quad (5.3)$$

Applying the geometrical method of the previous Section, two choices of the Lagrangian are as follows:

$$L_0 = \sum_{i=1}^N \left[ \frac{1}{2} (P_i \dot{Q}_i - Q_i \dot{P}_i) - \frac{1}{2} P_i^2 + e^{-(Q_{i+1} - Q_i)} \right]; \quad (5.4)$$

$$L = \sum_{i=1}^N \left[ \frac{1}{2} (P_i^2 + e^{-(Q_{i+1} - Q_i)}) \dot{Q}_i + \pi_i(P) \dot{P}_i \right] - H(Q, P) \quad (5.5)$$

where

$$\pi_i(P) = \frac{1}{2} \sum_{j=1}^N \epsilon(i - j) \dot{P}_j; \quad (5.6)$$

$$H(Q, P) = \sum_{i=1}^N \left[ \frac{P_i^3}{3} + (P_i + P_{i+1}) e^{-(Q_{i+1} - Q_i)} \right] \quad (5.7)$$

$f_{\mu\nu}$  turns out to have the canonical Poisson bracket structure

$$f_{\mu\nu} = \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix} \quad (5.8)$$

so that

$$f^{\mu\nu} = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \quad (5.9)$$

$F_{\mu\nu}$  can be shown to have the form [22]

$$F_{\mu\nu} = \begin{pmatrix} A & B \\ B & e \end{pmatrix} \quad (5.10)$$

where

$$\begin{aligned} A_{ij} &= \delta_{i+1,j} e^{-(Q_{i+1} - Q_i)} - \delta_{i,j+1} e^{-(Q_{j+1} - Q_j)} \\ B_{ij} &= P_i \delta_{ij}; \quad e_{ij} = \epsilon(j - i) \end{aligned} \quad (5.11)$$

The  $(1, 1)$  tensor  $S_\mu^\nu$  thus takes the form

$$S_\mu^\nu = \begin{pmatrix} B & A \\ -e & B \end{pmatrix} \quad (5.12)$$

and the conserved quantities are

$$Tr S = 2Tr B = 2 \sum_{i=1}^N P_i; \quad (5.13)$$

$$\frac{1}{2} Tr S^2 = Tr [2B^2 - (Ae + eA)] = \sum_{i=1}^N \left[ \frac{P_i^2}{2} + e^{-(Q_{i+1} - Q_i)} \right] \equiv H_0(Q, P); \quad (5.14)$$

$$\frac{1}{6} Tr S^3 = \sum_{i=1}^N \left[ \frac{P_i^3}{3} + (P_i + P_{i+1}) e^{-(Q_{i+1} - Q_i)} \right] \equiv H(Q, P) \quad (5.15)$$

The Lax representation (4.23) for the Toda equation takes the form of the following matrix equations

$$\frac{dA}{dt} = -[B, D]; \quad (5.16)$$

$$\frac{dB}{dt} = A - De = \frac{1}{2}[e, D] \quad (5.17)$$

which reduce to the Toda equations  $\dot{Q}_i = P_i$  and  $\dot{P}_i = e^{-(Q_i - Q_{i-1})} - e^{-(Q_{i+1} - Q_i)}$  respectively.

## 5.1 Group Structure of Toda Equations

Eq.(5.1) can be differentiated and put in the form

$$\begin{aligned} \ddot{Q}_1 &= -e^{-(Q_2 - Q_1)} \\ \ddot{Q}_i &= \dot{P}_i = e^{-(Q_i - Q_{i-1})} - e^{-(Q_{i+1} - Q_i)} \\ \ddot{Q}_N &= \dot{P}_N = e^{-(Q_N - Q_{N-1})} \end{aligned} \quad (5.18)$$

It is easily checked that

$$\sum_{i=1}^N \ddot{Q}_i = \sum_{i=1}^N \dot{P}_i = 0 \quad (5.19)$$

i.e., the total momentum is conserved, and therefore the centre of mass motion can be separated and the dynamics of the system expressed in terms of  $(N-1)$  coordinates and momenta. Defining

$$q_a = Q_{a+1} - Q_a; \quad a = 1, 2, \dots, N-1, \quad (5.20)$$

the second order equations satisfied by the  $q_a$ 's can be written as

$$\begin{aligned} \ddot{q}_1 &= 2e^{-q_1} - e^{-q_2} \\ \ddot{q}_a &= -e^{-q_{a-1}} + 2e^{-q_a} - e^{-q_{a+1}}; \quad a = 1, \dots, N-1 \\ \ddot{q}_N &= -e^{-q_{N-1}} + 2e^{-q_N} \end{aligned} \quad (5.21)$$

which can be compactly written as

$$\ddot{q}_a = \sum_{b=1}^{N-1} K_{ab} e^{-q_b} \quad (5.22)$$

$K_{ab}$  being the Cartan matrix for  $SU(N)$ . Eq.(5.22) generalizes for the other Lie Algebras as well.

The Lagrangian giving rise to the above Euler-Lagrangian equations can be written as

$$L = \sum_{a=1}^N \sum_{b=1}^N \frac{1}{2} \dot{q}_a K_{ab}^{-1} \dot{q}_b - \sum_{a=1}^N e^{-q_a} \quad (5.23)$$

$K_{ab}^{-1}$  being the inverse of the Cartan matrix. The momenta conjugate to  $q_a$  are defined as

$$p_a = \frac{\partial L}{\partial \dot{q}_a} = \sum_{b=1}^{N-1} K_{ab}^{-1} \dot{q}_b \quad (5.24)$$

and it is easily checked that

$$\{q_a, p_b\} = \delta_{ab} \quad (5.25)$$

so that  $\{q_a, p_a\}$  constitute a canonical coordinate system.

That the group structure entering above is not just accidental, can be seen by defining the following Lax operators:

$$S = \frac{1}{2} \sum_{a=1}^N [p_a H_a + (E_a + E_{-a}) e^{-q_a/2}]; \quad (5.26)$$

$$U = -\frac{1}{2} \sum_{a=1}^{N-1} e^{-q_a/2} [E_a - E_{-a}] \quad (5.27)$$

where  $H_a$  and  $E_a$  are the generators of  $SU(N)$  in the Chevally basis.

The Lax equation (4.23) can be seen to be satisfied, since  $\frac{dS}{dt} - [S, U]$  reduces to

$$\frac{1}{2} \sum_{a,b=1}^{N-1} H_a K_{ab}^{-1} [\ddot{q}_b - \sum_{c=1}^{N-1} K_{bc} e^{-q_c}]$$

which is zero by virtue of the Toda equations (5.22). Hence the quantities

$$K_n = \frac{1}{n} \text{Tr} S^n \quad (5.28)$$

must be conserved under the flow of the Toda equations. Since  $S$  belongs to the  $SU(N)$  algebra, the number of independent conserved quantities can equal  $(N-1)$ , which is the rank of  $SU(N)$ . The total number of conserved quantities is thus  $N$ , if we add the total momentum. It can be shown that these are also in involution [22]. This treatment is due to Leznov and Saveliev [18].

## 6 Zakharov-Shabat Formalism

So far we have only studied two integrable models, viz., the continuum  $KdV$  and the finite dimensional Toda Lattice. In trying to understand the non-linear Schroedinger equation which is also integrable, Zakharov and Shabat [23] obtained a description which was later generalized by AKNS [24] to describe various other integrable models. This approach uses a Lax operator which is first order in the derivative  $\partial_x$ , in contrast to the second order formalism in eq.(3.19). Besides describing various integrable models in a unified manner, this approach has the additional advantage that the inverse scattering method generalizes readily to the quantum case. In what follows, we describe the first order formulation of the Lax operator, and elucidate the essential features of this approach. It is easily checked that if

$$L(t)\psi(t) = -\lambda\psi(t); \quad (6.1)$$

$$\partial_t L(t) = [B(t), L(t)], \quad (6.2)$$

where

$$\frac{\partial\psi(t)}{\partial t} = B(t)\psi(t), \quad (6.3)$$

then

$$\frac{\partial\lambda(t)}{\partial t} = 0. \quad (6.4)$$

We can invert the argument to identify the Lax pair in the following way. Namely, if

$$L(t)\psi(t) = -\lambda\psi(t); \frac{\partial\psi(t)}{\partial t} = B(t)\psi(t), \text{ with } \frac{\partial\lambda(t)}{\partial t} = 0$$

, i.e., if the compatibility condition of (6.1) and (6.3) yield the system under study, then  $L(t)$  and  $B(t)$  can be identified as the Lax pair of the system. We would like  $L(t)$  to be linear in  $\partial_x$ . Using the analogy between the Klein-Gordon and Dirac equations, we define a two-component column matrix

$$\phi = \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} \quad (6.5)$$

and generalize the two equations (6.1) and (6.3) to first order matrix equations as

$$\begin{aligned} \frac{\partial\phi}{\partial x} &= (q\sigma_+ + r\sigma_- - i\rho\sigma_3)\phi; \\ \frac{\partial\phi}{\partial t} &= (P\sigma_+ + Q\sigma_- + R\sigma_3)\phi \end{aligned} \quad (6.6)$$

where  $\sigma_{\pm}$  and  $\sigma_3$  are the Pauli spin matrices. The dynamical variables  $q(x, t)$  and  $r(x, t)$  do not depend on the spectral parameter  $\rho$  which is assumed to be independent of  $x$  and  $t$ . The coefficient functions  $P$ ,  $Q$  and  $R$  on the other hand, do depend on  $\rho$ , and are functionals of  $q$  and  $r$ . Demanding that the partial derivatives of  $\phi$  w.r.t.  $x$  and  $t$  commute, we obtain the compatibility conditions to be

$$\frac{\partial R}{\partial x} = qQ - rP; \quad (6.7)$$

$$\frac{\partial r}{\partial t} = \frac{\partial Q}{\partial x} - 2rR - 2i\rho Q; \quad (6.8)$$

$$\frac{\partial q}{\partial t} = \frac{\partial P}{\partial x} + 2qR + 2i\rho P \quad (6.9)$$

i.e., if (6.7-9) describe the non-linear evolution of a system, then (6.6) describes the Lax pair appropriate for such a system. Explicitly

$$L = \partial_{\lambda} - q\sigma_+ - r\sigma_- \quad (6.10)$$

$$B = P\sigma_+ + Q\sigma_- + R\sigma_3 \quad (6.11)$$

so that (6.2) is satisfied.

The choice of  $r = 6$  yields the  $KdV$  equation, and the choice  $r - q = -iv/\sqrt{6}$ , the  $MKdV$  equation. The choice  $q = \sqrt{k}\psi^*$  and  $r = \sqrt{k}\psi$ ,  $k$  being an arbitrary constant parameter, yields the non-linear Schroedinger equation:

$$i\partial_t\psi = -\psi_{xx} + 2k|\psi|^2\psi \quad (6.12)$$

and the choice  $r = -q = \frac{1}{2}\omega_x$  with

$$P = Q = \frac{i}{4\rho}\sin\omega$$

yields the sine-Gordon equation.

The operator  $(L + \lambda)$  in (6.1) can be rewritten as  $v(x, t, \lambda) + \partial_x$ , where

$$v = -q\sigma_+ - r\sigma_3 + i\rho\sigma_3 \quad (6.13)$$

If one knows the solution of the associated Schroedinger equation at some other point  $(x, t)$  by multiplying the solution by a hermitian matrix  $T(x, y, t, \lambda)$ , i.e.,

$$\psi(x, t, \lambda) = T(x, y, t, \lambda)\psi(y, t, \lambda) \quad (6.14)$$

where  $T(x, y, t, \lambda)$  is a solution of

$$\partial_x T(x, y, t, \lambda) = -(q\sigma_+ - r\sigma_- + i\rho\sigma_3)T(x, y, t, \lambda) \quad (6.15)$$

with the initial condition  $T(x, x, t, \lambda) = I$ .

## 7 The Zero Curvature Condition

The Lax condition (6.2) can be written as

$$[\partial_t - B, L] = 0 \quad (7.1)$$

Using

$$L = \partial_x - A(x) \quad (7.2)$$

we obtain the form

$$[(\partial_t - B), (\partial_x - A)] = 0 \quad (7.3)$$

which is like a zero-curvature condition for

$$F_{01} = [(\partial_0 - A_0), (\partial_1 - A_1)] \quad (7.4)$$

with the identification

$$A_0 = -B(x, \rho); A_1 = -A(x, \rho) \quad (7.5)$$

The importance of the zero curvature condition stems from the fact that (6.6) may be solved, using

$$\psi(x) = T(x, y, \rho)\psi(y) \quad (7.6)$$

where the transformation

$$T(x, y, \rho) = P_r \exp\left[-\int_y^x A_1(z) dz\right] \quad (7.7)$$

where  $P_r$  denotes path ordering.

It is easy to see that  $T(x, y, \rho)$  translates solutions of the problem along the  $x$ -axis for a fixed time, i.e.,

$$T(x, y, \rho)T(y, z, \rho) = T(x, z, \rho); \quad (7.8)$$

$$T^{-1}(x, y, \rho) = T(y, x, \rho); \quad (7.9)$$

$$T(x, x, \rho) = 1 \quad (7.10)$$

Setting

$$U_r(x_2, t_2; x_1, t_1) = P_r \exp\left[-\int_{x_1, t_1}^{x_2, t_2} A_\mu dx^\mu\right] \quad (7.11)$$

and taking the product of two such exponents, it is easy to see that

$$U_{r_1}(x_2, t_2; x_1, t_1)U_{r_2}(x_1, t_1; x_2, t_2) = \exp\left[-\frac{1}{2} \oint_C d\sigma^{\mu\nu} F_{\mu\nu}\right], \quad (7.12)$$

using the Baker-Campbell-Hausdorff formula and the Stokes theorem, the integration being done over the area enclosed by the closed path  $r_1 + r_2$ . As the curvature  $F_{\mu\nu}$  vanishes,

$$U_{r_1}(x_2, t_2; x_1, t_1)U_{r_2}(x_1, t_1; x_2, t_2) = 1 \quad (7.13)$$

and so

$$U_r^{-1}(x_2, t_2; x_1, t_1) = U_r(x_1, t_1; x_2, t_2) \quad (7.14)$$

so that

$$U_{r_1}(x_2, t_2; x_1, t_1) = U_{r_2}(x_2, t_2; x_1, t_1) \quad (7.15)$$

ergo,  $U$  is independent of the path taken. For a closed path,  $U(x, t; x, t) = 1$ . Hence path ordering drops out of the transition matrix  $T(x, y, \rho)$ .

Returning to the time evolution of the transition matrix, it can be shown that

$$\partial_t T(x, y, \rho) = [B(x, \rho), T(x, y, \rho)] \quad (7.16)$$

which is the form of a Lax equation, so that all quantities of the form

$$K_n = \frac{1}{n} \text{Tr}[T(\rho)]^n; \quad K_0 = \ln[\det T(\rho)] \quad (7.17)$$

are conserved. We thus have an infinite number of conserved quantities when the zero curvature conditions are fulfilled. That this holds also for Toda Field Theories was shown by Olive and Turok [16].

## 8 From Conformal Invariance To Toda Field Theory

That the  $KdV$  equation has a hidden conformal symmetry can be seen by making a Fourier expansion with Fourier coefficients

$$u_n = -\frac{1}{4} \int_0^{2\pi} u(x) e^{-inx} \frac{dx}{2\pi} \delta_{n0} \quad (8.1)$$

It can be shown that the Poisson brackets of the  $u_n$  satisfy the Virasoro Algebra (up to trivial factors), i.e.,

$$-2i\pi\{u_n, u_m\} = -(n-m)u_{n+m} + \frac{1}{2}n(n^2-1)\delta_{n+m} \quad (8.2)$$

Higher order terms in the  $KdV$  hierarchy have a hidden  $\omega$  symmetry.

We now digress to take a look at Toda Field Theories. These are essentially the only class of integrable, interacting, conformally invariant field theories in two space-time dimensions. To see this, we start with the generic action

$$S = \int [\frac{1}{2} \partial_\mu \phi_i \partial^\mu \phi_i - V(\phi_i)] d^2 z \quad (8.3)$$

The trace of the naive conserved energy-momentum tensor becomes

$$T^\mu_\mu = 2V. \quad (8.4)$$

As the trace of the energy-momentum tensor is required to vanish in a conformally invariant theory, it seems that if  $V \neq 0$ , the theory is not conformal. However there is an ambiguity in the definition of the energy-momentum tensor. If we attempt to improve the naive energy-momentum tensor without violating the conservation property, we could choose

$$\theta_{\mu\nu} = T_{\mu\nu} + [\partial_\mu \partial_\nu - \eta_{\mu\nu} \partial^2] f(\phi_i) \quad (8.5)$$

whence the trace of the modified energy-momentum tensor is

$$\theta_{\mu\mu} = 2V + \partial_+ \partial_- f \quad (8.6)$$

$\pm$  being the light cone directions. If the second term is to cancel the first, we somehow need to get rid of the derivatives. This can be done, using the equations of motion. Without knowing the explicit equations of motion, the most general expression for  $f(\phi_i)$  is  $\sum c_i \phi_i$ . Using the equations of motion resulting from varying the action, the tracelessness condition becomes

$$2V + \sum_i c_i \frac{\partial V}{\partial \phi_i} = 0 \quad (8.7)$$

Eq.(8,7) is easily solved, with the result that the trace of the energy-momentum tensor vanishes if the potential is of the form

$$V(\phi_i) = \sum_j d_j \exp[\sum b_{ij} \phi_{ji}], \quad (8.8)$$

satisfying the requirement

$$\sum_i c_i b_{ij} = -2 \quad (8.9)$$

We choose  $b_{ij}$  to be related to the Cartan matrix of a simple Lie Algebra. The resulting field theories are called Toda Field Theories, and are described by the action

$$S_{Toda} = \int [\frac{1}{2} (\partial_\mu \phi, \partial^\mu \phi) - \frac{m^2}{\beta^2} \sum \exp(\beta \langle \alpha^{(i)}, \phi \rangle)] d^2 x \quad (8.10)$$

where  $\langle, \rangle$  is the scalar product in the root space, and  $\phi$  takes its values in the root space of the simple Lie Algebra on hand.

The equations of motion obtained from (8.10) are

$$\beta \partial^\mu \partial_\mu \phi_i + m^2 \exp(\sum K_{ij} \phi_j) = 0 \quad (8.11)$$

Specializing for the  $SU(n)$  group, and setting  $m = \beta = 1$ , this becomes

$$\partial_+ \partial_- \phi_i = -\exp(K_{ij} \phi_j) \quad (8.12)$$

With  $\phi_0 = 0$  and  $\phi_{i+1} = 0$ , this reduces to

$$\partial_+ \partial_- \phi_i = -\exp(2\phi_i - \phi_{i-1} - \phi_{i+1}) \quad (8.13)$$

Setting

$$\psi_i = (\phi_i - \phi_{i-1}) - (\phi_{i+1} - \phi_i),$$

after Mikhailov [27], we get the equation

$$\partial_t^2 \psi_i - \partial_x^2 \psi_i = -[2e^{\psi_i} - e^{\psi_{i-1}} - e^{\psi_{i+1}}] \quad (8.14)$$

which is easily seen to be related to the Toda equations (5.20). One expects that the Toda Field Theories are integrable, and it turns out that they are indeed so (see ref.[8]). The calculation rests upon the existence of a zero curvature condition for certain group theoretical combinations of  $\phi$ , which can be chosen as gauge fields.

As mentioned earlier, the Toda Field Theories have been completely solved for simple  $g$  by Leonov and Saveliev [18]. They have also been solved for affine  $g$  by Olive and Turok [16].

Quantization of the Toda Field Theories is more problematic since the potential has no local minimum, the latter being attained at infinity, using the gauge group  $A_1$ . A lucid discussion of the problems encountered in the theory is given in ref.[29].

The central charge of the Toda theories can be constructed using free field technology, and is found to be [30]

$$C = \frac{\hbar r}{2\pi} + 12 \left[ \frac{\hbar \beta \rho}{4\pi} + \frac{\rho^\vee}{\beta} \right]^2, \quad (8.15)$$

$r$  being the rank of the algebra,  $\rho$  being half the sum of the positive roots, and  $\rho^\vee$  its dual. Eq.(8.15) gives an indication that a quantum Toda theory with a strong coupling constant is equivalent to another Toda theory with a weak coupling constant, obtained by replacing  $\beta$  by  $4\pi/\hbar\beta$ , and interchanging roots and "coroots".

Incidentally, strong/weak coupling duality has recently become a subject of immense study in relation to string theories.

It is possible to obtain the minimal models from the Toda Field Theories. For a particular value of  $\beta$ , the central charges can be made to agree. However this is not enough. A complication arises from the fact that not all primary fields in the minimal models are actually present in the Toda theory. However, because of the duality in the theory, we can add another part of the potential with the coupling constant replaced by its dual; see Mansfield [31]. This modification is sufficient to give complete agreement.

## 9 W-Algebras: Hamiltonian Reduction of WZNW

Another fact which makes the conformally invariant Toda theories interesting is that to each such Toda theory, there corresponds a  $W$ -algebra. The  $W$ -algebras are an extension of the Virasoro algebra by adding primary fields of spin higher than  $Z$ , and were introduced by Zamolodchikov [32] as a pointer to conformal field theories with a larger overall symmetry. Zamolodchikov [32] investigated the case in which a primary field  $w(r)$  of weight 3 is added to the

Virasoro algebra. In order for the algebra to be close, it had to be made ‘non-linear’, and hence lost its linear Lie Algebra character.

Balog et al [33-35] showed that the Liouville and Toda Field Theories can be obtained as conformally reduced  $WZNW$  theories. This reduction can be viewed as a gauge procedure, and the Toda field theory can be obtained as the gauge invariant content of a gauged  $WZNW$  theory. The Liouville theory is obtained for the special case of the  $SL(2, R)$  gauge group.

The most powerful method of constructing  $W$ -algebras is through the so-called quantum Drinfeld-Sokolov reduction. In this, one starts with an affine Lie Algebra, and reduces it by imposing some constraint on its generators. At the classical level, this procedure which leads to the so-called Gelfand-Dickey algebras [36], was pioneered by Drinfeld and Sokolov [37].

It is thus clear that under the reduction that takes a  $WZNW$  field theory to a Toda field theory, the affine Lie Algebra that characterizes the  $WZNW$  theory reduces to a  $W$ -algebra that is associated to a Toda field theory. This approach is also readily generalizable to the supersymmetric case where various new  $W$ -superalgebras have been found as symmetry algebras of supersymmetric Toda field theories. We refer the interested reader to ref.[38] for further progress in this area.

In what follows, we review the essential steps of the Lagrangian reduction of the  $WZNW$  model. The  $WZNW$  action for a non-compact group  $G$  in 2D Minkowski space-time is

$$S(g) = -\frac{k}{8\pi} \int_{S^2} d^2\rho \eta^{\mu\nu} Tr(g^{-1}\partial_\mu g)(g^{-1}\partial_\nu g) + \frac{k}{12\pi} \int_B Tr(g^{-1}dg)^3 \quad (9.1)$$

where  $B$  is the volume occupied by  $S^2$ . The left and right Affine Kac-Moody [AKM] symmetries of this theory are generated by the Noether currents

$$J(\lambda) = \kappa Tr[\lambda(\partial_+ g)g^{-1}]; \quad \tilde{J}(\lambda) = -\kappa Tr[\lambda g^{-1}(\partial_- g)] \quad (9.2)$$

where  $\kappa = \frac{-k}{4\pi}$ , and  $\lambda$  is an element of the Lie Algebra  $\mathfrak{g}$ . The  $WZNW$  equations of motion are known to be equivalent to the current conservation

$$\partial_- J = \partial_+ \tilde{J} = 0. \quad (9.3)$$

We now choose the following Gauss decomposition of an arbitrary element  $g=ABC$ , e.g.,

$$\begin{aligned} A &= \exp\left[\sum_{\alpha \in \Delta^+} x^\alpha E_\alpha\right]; \\ B &= \exp\left[\frac{1}{2}\left(\sum_{\alpha \in \Delta} \phi^\alpha H_\alpha\right)\right]; \\ C &= \exp\left[\sum_{\alpha \in \Delta^-} y^\alpha E_\alpha\right]; \end{aligned} \quad (9.4)$$

where Cartan-Weyl root vectors  $E_\alpha$ , Cartan subalgebra generators  $H_\alpha = [E_\alpha, E_{-\alpha}]$ , and a set of positive (negative) roots  $\Delta^\pm$  have been introduced with the following properties

$$K_{\alpha\beta} = \alpha(H_\beta) = \frac{2\alpha \cdot \beta}{|\alpha|^2}; \quad \alpha, \beta \in \Delta; \quad |\alpha_{long}|^2 = 2; \quad (9.5)$$

$$Tr(H_\alpha \dot{H}_\beta) = \frac{2}{|\alpha|^2} K_{\alpha\beta} \equiv C_{\alpha\beta}; \quad (9.6)$$

$$Tr(E_\alpha \dot{E}_\beta) = \frac{2}{|\alpha|^2} \delta_{\alpha, -\beta}; \quad Tr[E_\alpha, H_\beta] = 0. \quad (9.7)$$

We also introduce the Polyakov-Wiegmann identity

$$\begin{aligned} S(ABC) &= S(A) + S(B) + S(C) + \kappa \int d^2\rho Tr[(A^{-1}\partial_- A)\partial_+ B]B^{-1} \\ &\quad + (B^{-1}\partial_- B)(\partial_+ C)C^{-1} + (A^{-1}\partial_- A)(B(\partial_+ C)C^{-1}B^{-1}) \end{aligned} \quad (9.8)$$



We now see, using eqs.(9.4-9.8), that the generalized constraints

$$J(E_\alpha) = \kappa c_1^\alpha; \quad \bar{J}(E_{-\alpha}) = -\kappa c_2^\alpha; \quad \alpha \in \Delta^+ \quad (9.9)$$

with some real numbers  $c_{1,2}^\alpha$  whose values do not vanish only for primitive roots  $\alpha \in \Delta$ , are enough to reduce the  $G$ -based WZNW theory to the Toda Field Theory defined by the Lagrangian

$$L_{Toda} = -\frac{k}{8\pi} \left[ \frac{1}{4} C_{\alpha\beta} \partial_+ \phi^\alpha \partial_- \phi^\beta - \sum_{\alpha \in \Delta} (u^2)^\alpha e^{\left(\frac{1}{2} K_{\alpha\beta} \phi^\beta\right)} \right] \quad (9.10)$$

where

$$(u^2)^\alpha = |\alpha|^2 c_1^\alpha c_2^\alpha$$

. Due to  $c_{1,2}^\alpha \neq 0$  for the primitive roots, the constraint (9.7) can be re-written in terms of the Gauss decomposition (9.5-7) as follows:

$$\begin{aligned} A^{-1} \partial_- A &= B \left[ \sum_{\alpha \in \Delta} \frac{1}{2} |\alpha|^2 c_2^\alpha E_\alpha \right] B^{-1} \\ &= \sum_{\alpha \in \Delta} \frac{1}{2} |\alpha|^2 c_2^\alpha E_\alpha \exp \left[ \frac{1}{2} K_{\alpha\beta} \phi^\beta \right]; \end{aligned} \quad (9.11)$$

$$\begin{aligned} (\partial_+ C) C^{-1} &= B^{-1} \left[ \sum_{\alpha \in \Delta} \frac{1}{2} |\alpha|^2 c_1^\alpha E_{-\alpha} \right] B \\ &= \sum_{\alpha \in \Delta} \frac{1}{2} |\alpha|^2 c_1^\alpha E_{-\alpha} \exp \left[ \frac{1}{2} K_{\alpha\beta} \phi^\beta \right]; \end{aligned} \quad (9.12)$$

In the WZNW equations of motion,  $A$  and  $C$  occur only in the combinations given in (9.11-12), so that they can be eliminated in favour of  $B$  or  $\phi^\alpha$ . The remaining equation is just the Toda equation [25,34,35]:

$$\partial_+ \partial_- \phi^\alpha + \frac{1}{2} |\alpha|^2 (u^\alpha)^2 \exp \left[ \frac{1}{2} K_{\alpha\beta} \phi^\beta \right] = 0; \quad (9.13)$$

(see also ref.[25] for details).

As mentioned earlier, the Toda Field Theory possesses an extended symmetry represented by a classical  $W$ -algebra. These  $W$ -algebras can be obtained as the quantum versions of the so-called Gelfand-Dickey algebras [36] known in the theory of  $KdV$  equations. For instance, the Poisson bracket associated with the  $KdV$  equation in (8.2), results in the classical version of the Virasoro algebra which is the simplest  $W$ -algebra. Moreover, the Lax representation of the  $KdV$  equation (3.15), defines the third order differential operator  $B = w^{(3)}$ . The Fourier components of  $B$ , along with those of the  $KdV$  field, form the Gelfand-Dickey [36] algebra that generalizes to  $w^{(3)}$  in the quantum case.

Now regarding the Toda theory as a constrained WZNW theory, the Hamiltonian structure can be obtained by a classical Drinfeld-Sokolov reduction from the constrained phase space of the  $AKM$  algebra. In the Hamiltonian formalism, the  $AKM$  symmetry of the WZNW theory is represented by first class constraints. The  $W$ -algebra of the Toda theory arises as the Poisson bracket algebra of gauge-invariant polynomials of the constrained  $AKM$  currents and their derivatives. In what follows, we summarize the arguments supporting these statements.

Let  $g(z, \bar{z})$  be the  $G$ -valued WZNW fields and  $J(z)$  the corresponding  $AKM$  currents having the form

$$g(z, \bar{z}) = g(z)g(\bar{z}); \quad \partial g(z) = J(z)g(z) \quad (9.14)$$

Let  $\dim g$  be the dimension of  $G$ ;  $l$  its rank;  $k$  the level of the associated  $AKM$  algebra  $\hat{g}$ ;  $g$  the dual Coxeter number of  $G$ ;  $\rho$  the half sum of the positive roots; and  $\beta$  the dual of  $\rho$ .

The constrained  $WZNW$  theory is specified by (9.9). After a suitable choice of constants  $c_i$ , the currents  $J(z)$  can be decomposed as

$$\begin{aligned} J(z) &= I_- + j(z); \quad I_- = \sum_{i=1}^l E_{-\alpha_i}; \\ j(z) &= \sum_{i=1}^l j^i(z) H_i + \sum_{\phi \in \Delta^+} E_\phi \end{aligned} \quad (9.15)$$

where  $\{E_{\alpha_i}\}$  are  $l$  simple roots of  $g$ . The maximal subgroup of  $\hat{G}$  leaving this form of currents invariant, is the maximal nil-potent subgroup generated by  $E_\phi$ , ( $\phi \in \Delta^+$ ), and implemented by the  $(\dim g - l)/2$  constrained  $AKM$  currents  $J^\phi(z)$ . This allows us to interpret the constrained  $WZNW$  theory as the gauge theory in which all but  $l$  of the  $(\dim g + l)/2$  components of  $J$  are gauge components [33-35].

The current  $j(z)$  and the gauge transformations corresponding to  $E_\phi$  act on each column of the  $WZNW$  field  $g(z)$  separately, while each column contains only one gauge-invariant component  $e$  (of the highest weight), satisfying  $E_\phi e = 0$ . The gauge degrees of freedom corresponding to the other elements of each column can be eliminated by a gauge fixing in favour of  $e$ . Because of (9.15), this leads to a linear pseudo-differential equation  $De = 0$ , where  $D$  is a polynomial pseudo-differential operator whose coefficients are gauge invariant polynomials in the currents  $J$ . This operator  $D$  can now be used to define a classical  $W$ -algebra by choosing a Drinfeld-Sokolov gauge in which one has

$$j_{DS} = \sum w^P(z) F_P \quad (9.16)$$

where  $P$ 's are the orders of  $l$  independent Casimir operators of  $g$ , and  $F_P$  generators with  $H$  weights  $(P - 1)$ , so that the gauge-fixed current (9.16) has only one non-vanishing component in each of the  $l$  irreducible representations in a decomposition of the adjoint of  $g$  w.r.t. one of its sub-groups  $SL(2, R)$ . The Poisson brackets between the different polynomials  $w^P$  define a classical  $W$ -algebra.

We close this Section by noting that Toda field theories also play an important role in the discussion of  $W$ -gravity, where they arise as effective quantum theories [39,40] for the  $W$ -gravity degrees of freedom in the conformal gauge. For a quantum version of the  $WZNW \rightarrow Toda$  conformal reduction, see [34, 41].

## 10 Self-Dual Y-M Theories: 2D Integrable Models

The self-dual Yang-Mills (SDYM) theory appears to be a master theory for a whole variety of 2D integrable systems, as we are now going to explain. Though there is no general proof, the statement can be checked on a case by case basis. The main point is that the 4D self-duality condition admits of a zero curvature representation underlying a Hamiltonian description of SDYM descendents in lower dimensions. This makes it possible to apply the inverse scattering method for integration of the SDYM equations. Simultaneously, it explains the origin of gauge symmetries in integrable systems of the  $KdV$  type, since the SDYM theory is both gauge and conformally invariant in 4D. And last but not least, this connection provides us with a systematic way to associate the  $KdV$  type hierarchy with any simple Lie Algebra.

SDYM solutions invariant by the action of a subgroup with two conformal generators satisfy a 2D differential equation, since each 1D subgroup reduced the number of independent variables by one. This allows us to describe the invariant SDYM solutions in terms of a 2D integrable system. All known 2D integrable systems seem to be derivable this way, by appropriate truncations of a 4D self-dual gauge theory. This is true, in particular, for the  $KdV$  and non-linear Schrödinger equations, the Liouville and Toda equations, as well as other integrable in 2 and 3 dimensions. Our presentation in this Section is only illustrative; we give one explicit example of embedding of the  $KdV$  equation into the 4D SDYM theory [42], and a supersymmetric generalization.

Let  $x^a = (x, y, z, t)$  be the coordinates of a flat 4D space-time of signature  $(+, +, -, -)$ . The invariant metric reads

$$ds^2 = 2dx dz - 2dy dt \quad (10.1)$$

The SDYM equations in 2+2 dimensions ( $\epsilon_{xyzt} = 1$ ) read as

$$F_{ab} = \frac{1}{2} \epsilon_{abcd} F^{cd} \quad (10.2)$$

and are equivalently represented by 3 equations of the form

$$F_{tx} = F_{yz} = F_{ty} + F_{xz} = 0 \quad (10.3)$$

After a dimensional reduction which is equivalent to setting

$$\partial_y = \partial_z - \partial_x = 0, \quad (10.4)$$

(10.3) takes the form

$$[\partial_t - H, \partial_x - Q] = [P, B] = 0; \quad [H, B] = [\partial_x - Q, \partial_x - P] \quad (10.5)$$

where

$$A_t = H; \quad A_x = Q; \quad A_y = -B; \quad A_z = P$$

It is clear that the first equation in (10.5) is a zero curvature equation. We now choose the non-compact group and an embedding pattern in the form

$$B = \begin{pmatrix} 0 & 0 \\ I & 0 \end{pmatrix}; \quad (10.6)$$

$$Q = \begin{pmatrix} \lambda & 1 \\ -u & -\lambda \end{pmatrix} \quad (10.7)$$

where  $\lambda$  is a constant and  $u = u(t, x + z)$ . We can expand the Lie Algebra-valued fields  $H$  and  $P$  as

$$H = H_- \tau_+ + H_+ \tau_- + H_3 \tau_3; \quad P = P_- \tau_+ + P_+ \tau_- + P_3 \tau_3 \quad (10.8)$$

where  $\tau_{\pm} = (\tau_1 \pm i\tau_2)/2$ , and  $\tau_{1,2,3}$  are the Pauli spin matrices. It is clear that the second equation of (10.5) gives

$$P_- = P_3 = 0, \quad (10.9)$$

while the third equation of (10.5) gives

$$H_- = -P_+; \quad H_3 = -\frac{1}{2} \partial_x (u + P'_+) - \lambda P_+ \quad (10.10)$$

where primes denote derivatives w.r.t.  $x$ .

Finally, the first equation of (10.5) yields 3 equations

$$\begin{aligned} H_+ &= u P_+ - \lambda \partial_x P_+ - \frac{1}{2} \partial_x \partial_x (u + P_+); \quad \partial_x (u + 2P_+) = 0; \\ \dot{u} &= \frac{1}{2} \partial_x \partial_x \partial_x (u + P_+) + (u - P_+) \partial_x u + 2\lambda^2 P_+ \end{aligned} \quad (10.11)$$

It follows that

$$\begin{aligned} P_+ &= -\frac{1}{2} u; \quad H_+ = -\frac{1}{2} u^2 + \frac{\lambda}{2} u_x - \frac{1}{4} u_{xx}; \\ \dot{u} &= \frac{1}{4} u_{xxx} + \frac{3}{2} u u_x - \lambda^2 u_x \end{aligned} \quad (10.12)$$

Changing the notation as

$$u \rightarrow u + \frac{2}{3}\lambda^2; \quad t \rightarrow 4t; \quad x + y \rightarrow x,$$

one obtains the *KdV* equation

$$u_t = u_{xxx} + 6uu_x.$$

This example may be relevant towards an ultimate unification of 2D integrable models and 2D conformal field theories, as well as within the 4D SDYM theories which are also closely related to  $N + 2$  strings.

## 10.1 Self-Duality and Supersymmetry

Extended Supersymmetry is compatible with self-duality in 2+2 dimensions. Therefore the Supersymmetric self-dual Yang-Mills theory (SSDYM) is capable of generating Supersymmetric 2D integrable models. However a Supersymmetric generalization of the SDYM theory is not unique. One could either replace a gauge group by its graded version, or a 2+2 dimensional space-time by superspace.

Supersymmetric generalizations of the *KdV* equation in 1+1 dimensions were obtained independently by Manin and Radul [43], Mathieu [44], Bilal and Gervais [45]: These equations have two dynamical variables, one bosonic  $u(x, t)$ , and one fermionic  $\psi(x, t)$ , and read

$$\begin{aligned} \partial_t u &= \frac{1}{2}u_{xxx} + 3u\partial_x u + \frac{3}{2}(\psi_{xx})\psi \\ \partial_t \psi &= \frac{1}{2}\psi_{xxx} + \frac{3}{2}\partial_x(u\psi) \end{aligned} \quad (10.13)$$

They are invariant under the  $N = 1$  Supersymmetry transformations

$$\delta u = \epsilon \partial_x \psi; \quad \delta \psi = \epsilon u \quad (10.14)$$

$\epsilon$  being a constant Grassmann parameter. Eqs.(10.13) are integrable, and can be obtained from the zero curvature condition associated with the graded Lie Algebra  $osp(2, 1)$

$$\partial_t A_x - \partial_x A_t + [A_t, A_x] = 0 \quad (10.15)$$

when the following ansatz is used for 2D Yang-Mills potentials [45]:

$$2A_t(x, t) = \begin{pmatrix} u_x & u_{xx} + 2u^2 + \psi_x\psi & -i\psi_{xx} - 2iu\psi \\ -2u & -u_x & i\psi_x \\ i\psi_x & i\psi_{xx} + 2iu\psi & 0 \end{pmatrix} \quad (10.16)$$

$$A_x = \begin{pmatrix} 0 & u & -i\psi \\ -1 & 0 & 0 \\ 0 & i\psi & 0 \end{pmatrix} \quad (10.17)$$

The 2D Super *KdV* can be embedded into the self-duality equations by choosing the  $osp(2/1)$ -valued matrices  $H, Q$  as  $H = A_t(x, t)$ ,  $Q = A_x(x, t)$ , and  $B, P$  as  $3 \times 3$  matrices as

$$B = \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad (10.18)$$

$$P = \begin{pmatrix} 0 & \frac{u}{2} & -\frac{3i}{4}\psi \\ 0 & 0 & 0 \\ 0 & \frac{3i}{4}\psi & 0 \end{pmatrix} \quad (10.19)$$

It can also be shown that the  $N = 1$  and  $N = 2$  Super *KdV* equations, as well as the  $N = 1$  Super-Liouville and Super-Toda equations, can all be obtained from the  $N = 2$  SSDYM theory by dimensional reductions and truncations [46]. A detailed analysis is however outside the scope of this Article.

## 11 Conclusions

It has been our aim to present a bird's eye view of the important developments in Integrable Systems over the past few decades. What has been achieved is possibly a more subjective viewpoint, related to building connections between sundry topics of immediate interest. It has certainly not been possible to delve more deeply into the fascinating developments in affine Toda Field Theory which seems to be a thrust area of research today. We refer to the excellent lecture series by Corrigan [48] on this subject. Neither is it possible to present an account of the interesting link between the  $KdV$  theory and Matrix models, encompassing thus 2D gravity; (see ref.[25] for a readable account). Supersymmetric Toda Field Theories have also been given the go by. They were first studied by Evans and Hollowood [49], as well as by Leites et al [50]. It seems to be possible to construct Toda Field Theories based on Lie Superalgebras with one proviso, namely, that it is necessary that the Lie Superalgebra admits a purely fermionic root system. This is only possible for the following algebras:

$$A(n, n-1); B(n, n); B(n-1, n); D(n, n-1); D(n, n); \text{ and } D(2, 1, \alpha).$$

In the generic case,  $N = 1$  Supersymmetric theories are obtained, which can be formulated in  $N = 1$  superspace. There is one special case, namely, the  $sl(n, n-1)$  theories have in fact  $N = 2$  Supersymmetry; see [49]. Recently, Brink and Vasiliev [51] have proposed a model generalizing  $A_N$  Toda Field Theories based on a continuous parameter, such that when this parameter takes on certain discrete values, the model reduces to the ordinary  $A_N$  Toda Theories. More recently, Wyllard [52] has worked out a  $WZNW$  reduction of these generalized theories, and has also attempted a Supersymmetric generalization [53] of the same.

One could also picture the affine Toda theories as integrable deformations of the conformal Toda theories. As an example, by adding an extra simple root to the  $A_1$  Toda theory, one obtains the affine Toda theory, which is also the sinh-Gordon theory. General integrable deformations have been investigated by Zamolodchikov, among others, as an interesting field. Toda theories also appear in many other diverse areas of theoretical physics, e.g., 1D discrete versions appear in the physics of monopoles [54]. Further, certain 3D continuous Toda systems are relevant to the classification of hyper-Kähler metrics in 4D [55]. Finally, it also appears that Toda Field Theories are relevant to  $M$ -Theory, the conjectured all-in-all Theory encompassing all String Theories [56].

I am deeply grateful to Prof. S.K. Malik, for providing me an opportunity to write this topical Article on the subject of Toda Field Theories. The literature on this subject is quite vast, but apart from my own interest in Lie Algebras and super-algebras (which often draws me into this area because of their obvious relevance to this subject), I have benefitted greatly from some concentrated literature [22,25,56,57]. I am especially grateful to Prof. Ashok Das for his encouragement, and for sparing time to go through the manuscript. I would also like to thank Jens Fjelstad for help with LaTeX.

## References

- [1] J.Liouville, J.Math Puris et Applique'es **18**, 71 (1853).
- [2] A.M.Polyakov, Phys. Lett.**B103**, 207 (1981).
- [3] J.L.Gervais and A.Neveu, Nucl.Phys.**B199**, 59 (1982).
- [4] E.D'Hoker and R.Jackiw, Phys.Rev.**D26**, 3517 (1982).
- [5] T.L.Curtweight and C.B.Thorn, Phys.Rev.Lett.**48**, 1309 (1982).
- [6] D.J.Korteweg and G.de Vries, Phil.Mag.**39**, 422 (1895).
- [7] C.S.Gardner, J.M.Greene, J.M.Kruskal, R.M.Miura, Phys.Rev.Lett.**19**, 1095 (1967).
- [8] R.M.Miura, J.Math.Phys.**9**, 1202 (1968).

- [9] R.M.Miura, C.S.Gardner, J.M.Kruskal, *J.Math.Phys.***9**, 1204 (1968).
- [10] C.H.Su, C.S.Gardner, *J.Math.Phys.***10**, 536 (1969).
- [11] C.S.Gardner, *J.Math. Phys.***12**, 1548 (1971).
- [12] J.M.Kruskal et al, *J.Math.Phys.***11**, 952 (1970).
- [13] C.S.Gardner et al, *Comm Pure And Appl Maths* **27**, 97 (1974).
- [14] M.Toda, *J.Phys.Soc.Of Japan* **22**, 431 (1967).
- [15] R.Hirota, K.Suzuki, *J.Phys.Soc.Of Japan* **28**, 1336 (1970).
- [16] D.Olive, M.Turok, *Nucl.Phys.***B257**, 277 (1985).
- [17] T.Hollowood, *Nucl.Phys.***B384**, 523 (1992).
- [18] A.N.Leznov, M.V.Saveliev, *Lett.Math.Phys.***3**, 489 (1979).
- [19] T.S.Hollowood, P.Mansfield, *Phys.Lett.***B226**, 73 (1989).
- [20] I.M.Gelfand, B.M.Levitan, *Trans.Amer.Math.Soc.***1**, 253 (1951).
- [21] P.D.Lax, *Comm Pure And Appl Maths* **21**, 467 (1968).
- [22] Ashok Das, *Integrable Models*, World Scientific (1989).
- [23] V.E.Zakharov, A.B.Shabat, *Soviet Phys. JETP* **34**, 62 (1972).
- [24] M.J.Ablowitz et al, *Phys.Rev.Lett.***30**, 1262 (1973); **31**, 125 (1973).
- [25] S.Ketov, *Conformal Field Theory*, World Scientific (1995).
- [26] A.Lenard (Unpublished); reported in Ref.[1].
- [27] A.V.Mikhailov, *Soviet Phys. JETP Lett.***30**, 414 (1979).
- [28] A.N.Leznov, M.Saveliev, *Comm Math Phys.***74**, 111 (1980).
- [29] Y.Karama, H.Nicolai, *Int J of Mod Phys.***A9**, 667 (1994).
- [30] P.Mansfield, *Nucl.Phys.***B222**, 419 (1983).
- [31] P.Mansfield, *Phys.Lett.***B242**, 387 (1990).
- [32] A.B.Zamolodchikov, *Theor. Math. Phys.* **65**, 1205 (1986).
- [33] J.Balog et al, *Phys.Lett.***B227**, 214 (1989).
- [34] L. Feher et al, *Phys.Rep.***222C**, 1 (1992).
- [35] P.Forgacs et al, *Phys.Lett.***B237**, 214 (1989).
- [36] L.Gelfand And L.A.Dikin, in *Gelfand, Collected Papers*, Ed by Gindinkin et al, Springer Verlag, N.Y. (1987); p625.
- [37] I.Drinfield, A.Sokolov, *J Sov Math.***30**, 1975 (1984).
- [38] P.Bouwknegt and K.Schoutens, *Phys.Rep.***223C**, 183 (1993).
- [39] E.Bergshoeff, A.Bilal, K.S.Stelle, *Int J Mod Phys.***A6**, 4951 (1991).
- [40] K.Schoutens, A.Sevrin, P.van Nieuwenhuizen, *Nucl.Phys.***B349**, 791 (1991).

- [41] L.O.Raifeartaigh, P.Ruelle, I.tsutsui, Phys.Lett.**B258**, 359 (1991).
- [42] L.J.Mason, G.A.J.Sparling, Phys.Lett.**A137**, 29 (1989).
- [43] Yu.L Manin and A.O.Radul, Comm. Math. Phys.**98**, 67 (1985).
- [44] P.Mathieu, Phys.Lett.**B203**, 287 (1988).
- [45] A.Bilal, J.L.Gervais, Phys.Lett.**B211**, 85 (1988).
- [46] A.Das, C.A.P.Galvao, Mod.Phys.Lett.**A8**, 1399 (1993).
- [47] S.J.Gates, H.Nishino, Phys.Lett.**B299**, 255 (1993).
- [48] E.Corrigan LANL **hep-th/9412213**, December 1994.
- [49] J.Evans, T.Hollowood, Nucl.Phys.**B352**, 723 (1991).
- [50] D.A.Leites, M.V.Saveliev And V.V.Sergonov, in  
*Group Theoretical Methods In Physics*, Vol.1,  
eds M.A.Markov et al, VNU Science Press, 1986.
- [51] L.Brink, M.Vasiliev, Nucl.Phys.**B459**, 273 (1995).
- [52] N.Wyllard, Mod.Phys.Lett.**A12**, 95 (1997).
- [53] N.Wyllard, Nucl.Phys.**B491**, 461 (1997).
- [54] H.Lu, C.N.Pope, K.W.Xu, LANL **hep-th/9604058**, April 1996.
- [55] I.Bakas, K.Sfestos, LANL **hep-th/9604003**, April 1996.
- [56] N.Wyllard, Ph.D Thesis (May 1998), Chalmers U; Goteborg(1998).
- [57] E.G.B.Hohler, Ph.D thesis, Trondheim Univ.Press (1995)

# 17. Perspectives Of Light-Front Quantum Field Theory: Some New Results<sup>\*†</sup>

Prem P. Srivastava<sup>‡</sup>

*Stanford Linear Accelerator Center, Stanford University, Stanford,  
CA 94309, USA*

## Abstract

A review of some basic topics in the light-front (LF) quantization of relativistic field theory is made. It is argued that the LF quantization is *equally appropriate* as the conventional one and that they lead, assuming the microcausality principle, to the same physical content. This is confirmed in the studies on the LF of the spontaneous symmetry breaking (SSB), of the degenerate vacua in Schwinger model (SM) and Chiral SM (CSM), of the chiral boson theory, and of the QCD in covariant gauges among others. The discussion on the LF is more economical and more transparent than that found in the conventional equal-time quantized theory. The removal of the constraints on the LF phase space by following the Dirac method, in fact, results in a substantially reduced number of independent dynamical variables. Consequently, the descriptions of the physical Hilbert space and the vacuum structure, for example, become more tractable. In the context of the Dyson-Wick perturbation theory the relevant propagators in the *front form* theory are causal. The Wick rotation can then be performed to employ the Euclidean space integrals in momentum space. The lack of manifest covariance becomes tractable, and still more so if we employ, as discussed in the text, the Fourier transform of the fermionic field based on a special construction of the LF spinor. The fact that the hyperplanes  $x^\pm = 0$  constitute characteristic surfaces of the hyperbolic partial differential equation is found irrelevant in the quantized theory; it seems sufficient to quantize the theory on one of the characteristic hyperplanes.

---

<sup>\*</sup>Research partially supported by the Department of Energy under contract DE-AC03-76SF00515.

<sup>†</sup>Slac report: SLAC-PUB-8219, August 1999.

<sup>‡</sup>E-mail: (1) prem@lafexsu1.lafex.cbpf.br; (2) prem@slac.stanford.edu. On leave of absence from *Instituto de Física, UERJ- Universidade do Estado de Rio de Janeiro*, RJ, Brasil.



## Contents

<b>1</b>	<b>Introduction</b>	<b>439</b>
1.1	Light-Front Quantized Theory . . . . .	440
1.2	LF Poincaré and IMF Generators. LF Spin Operator . . . . .	441
<b>2</b>	<b>LF quantized scalar theory</b>	<b>441</b>
2.1	Covariant Phase Space Factor on the LF . . . . .	441
2.2	LF Commutator . . . . .	442
2.3	Length Dimensions $L_{\perp}$ and $L_{\parallel}$ . . . . .	442
2.4	LF Hamiltonian. Dirac Procedure . . . . .	443
2.5	Scalar Field Propagator in momentum space . . . . .	444
2.6	First class constraint. Symmetry in $x^+$ and $x^-$ . . . . .	445
<b>3</b>	<b>SSB Mechanism, Topological Kink Solution, and Chiral Boson theory on LF</b>	<b>445</b>
3.1	SSB in two dimensional scalar theory . . . . .	445
3.2	Spontaneously broken continuous symmetry . . . . .	447
3.3	Kink solution and Topological quantum number . . . . .	448
3.4	Chiral Boson theory on the LF . . . . .	448
<b>4</b>	<b>LF quantized Dirac field</b>	<b>451</b>
4.1	Anticommutators . . . . .	451
4.2	LF Spinor in momentum space and its properties . . . . .	452
4.3	Fermion propagator . . . . .	453
4.4	$\Gamma_5$ Symmetry. Chirality transformation on the LF . . . . .	454
4.5	Helicity Operator, LF Majorana and Weyl fermions . . . . .	456
4.6	Bilocal operators . . . . .	457
<b>5</b>	<b>LF quantization of Gauge theory</b>	<b>457</b>
<b>6</b>	<b>Vacuum Structures in Schwinger and Chiral Schwinger Models</b>	<b>459</b>
<b>7</b>	<b>QCD in Covariant gauges</b>	<b>463</b>
<b>8</b>	<b>Conclusions</b>	<b>465</b>

## 1 Introduction

Half a century ago, Dirac [1] discussed the unification, in a relativistic theory, of the principles of the quantization and the special relativity theory which were by then firmly established. The Light-Front (LF) quantization which studies the relativistic quantum dynamics of physical system on the hyperplanes :  $x^0 + x^3 \equiv \sqrt{2}x^+ = \text{const.}$ , called the *front form* theory, was also proposed and some of its advantages pointed out. The *instant form* or the conventional equal-time theory on the contrary uses the  $x^0 = \text{const.}$  hyperplanes. The LF coordinates  $x^\mu : (x^+, x^-, x^\perp)$ , where  $x^\pm = (x^0 \pm x^3)/\sqrt{2} = x_\mp$  and  $x^\perp = (x^1, x^2) = (-x_1, -x_2)$ , are convenient to use in the *front form* theory. They are *not related by a finite Lorentz transformation* to the coordinates  $(x^0 \equiv t, x^1, x^2, x^3)$  usually employed in the *instant form* theory and as such the descriptions of the same physical content in a dynamical theory on the LF, which studies the evolution of the system in  $x^+$  in place of  $x^0$ , may come out to be different from that given in the conventional treatment. This was found to be the case, for example, in the description of the spontaneous symmetry breaking (SSB) mechanism (Sec. 3) some time ago and in the recent study (Sec. 6) of some soluble two-dimensional gauge theory models, where it was also demonstrated that LF quantization is very economical in displaying the relevant degrees of freedom, leading directly to a physical Hilbert space. The LF quantized field theory may perhaps also be of some relevance in the understanding of the unification of the principles of the quantization with that of the general covariance [2].

We recall that the field theory at infinite momentum was employed in the context of the current algebra sum rules [3]. The Feynman rules adapted for infinite momentum frame (IMF), which were used by Weinberg [4], showed substantial simplifications in the context of the old fashioned perturbation theory computations. In the deep inelastic region with the IMF limit Bjorken [5] predicted the scaling of the deep inelastic structure functions. The parton model [6] of Feynman was also formulated in the IMF. At the same time the connection between the use of the LF variables and the IMF limit was being noticed by several authors [7], which prompted gradually the interest in the study of the *front form* theory as proposed by Dirac.

More recently, the interest in LF quantization has been revived [8, 9, 10] due to the difficulties encountered in the computation, in the conventional framework, of the nonperturbative effects in the context of QCD and the problem of the relativistic bound states of fermions [8, 9] in the presence of the complicated vacuum. Studies show [9, 8, 11] that the application of Light-front Tamm-Dancoff method may be feasible here. The technique of the regularization on the lattice has been quite successful for some problems but it cannot handle, for example, the bound states of light (chiral) fermions and has not been able yet to demonstrate, for example, the confinement of quarks. The problem of reconciling the standard constituent quark model and the QCD to describe the hadrons is also not satisfactorily resolved. In the former we employ few valence quarks while in the latter the QCD vacuum state itself contains, in the conventional theory, an infinite sea of constituent quarks and gluons (partons) with the density of low momentum constituents getting very large in view of the infrared slavery. The *front form* dynamics may serve as a complementary tool to study such problems, since we may possibly arrange to have a simple vacuum in it while transferring the complexity of the problem to the LF Hamiltonian. In the case of the scalar field theory, for example, the corresponding LF Hamiltonian is, in fact, found [12] to be *nonlocal* due to the presence of *constraints* on the *LF phase space*.

The LF quantization of QCD in its Hamiltonian form provides an alternative approach to lattice gauge theory for the computation of nonperturbative quantities, such as [8, 9] the spectrum and the LF Fock state wavefunctions of relativistic bound states. LF variables have found natural applications in several contexts, for example, in the quantization of (super-) string theory and M-theory [13]. They have also been employed in the nonabelian bosonization [14] of the field theory of  $N$  free Majorana fermions. The vacuum structures [15, 16] in the LF quantized Schwinger model (SM) and the Chiral SM (CSM) have been recently studied in a transparent fashion. The LF quantized QCD in covariant gauges has also been studied [17] in the context of the Dyson-Wick perturbation theory, where it is shown that the lack of manifest covariance in the calculations

becomes more tractable thanks to a useful form of the LF spinor introduced (Sec. 4). The relevant propagators are shown to be causal and the Wick rotation can be performed [18] to go over to the Euclidean space integrals allowing for the dimensional regularization to be used. The *front form* theory has also found applications in the nonperturbative sector of QCD in the context of the Bethe-Salpeter dynamics. The Covariant Instanteity ansatz (CIA) [19] introduced earlier, which invokes the Markov-Yukawa Transversality Principle, has been extended now to the covariant null plane (CNPA) [20, 21].

### 1.1 Light-Front Quantized Theory

We will make the *convention* to regard  $x^+ \equiv \tau$  as the LF-time coordinate while  $x^-$  as the *longitudinal spatial* coordinate. We note that  $[x^+, i\partial^-] = [x^-, i\partial^+] = -i$  where  $\partial^\pm = \partial_\mp = (\partial^0 \pm \partial^3)/\sqrt{2}$  etc. so that the coordinates  $x^+$  and  $x^-$  appear in a symmetric fashion. In terms of the null vector  $n^\mu = (1, 0, 0, 1)/\sqrt{2}$  and its dual  $\tilde{n} = (1, 0, 0, -1)/\sqrt{2}$ , with  $n \cdot n = \tilde{n} \cdot \tilde{n} = 0$ ,  $\tilde{n} \cdot n = 1$ , they may be written also as  $x^+ = n \cdot x$  and  $x^- = \tilde{n} \cdot x$  (See also Sec. 5). The temporal evolution in  $x^0$  or  $x^+$  of the system is generated by the Hamiltonians which are different in the two *forms* of the theory.

Consider [16, 10] the invariant distance between two spacetime points :  $(x - y)^2 = (x^0 - y^0)^2 - (\vec{x} - \vec{y})^2 = 2(x^+ - y^+)(x^- - y^-) - (x^\perp - y^\perp)^2$ . On an equal  $x^0 = y^0 = \text{const.}$  hyperplane the points have spacelike separation except for if they are *coincident* when it becomes lightlike one. On the LF with  $x^+ = y^+ = \text{const.}$  the distance becomes *independent of*  $(x^- - y^-)$  and the separation is again spacelike; it becomes lightlike one when  $x^\perp = y^\perp$  but with the difference that now the points need *not* necessarily be coincident along the longitudinal direction. The LF field theory hence *need not necessarily be local in*  $x^-$ , even if the corresponding *instant form* theory is formulated as a local one. For example, the commutator  $[A(x^+, x^-, x^\perp), B(0, 0, 0^\perp)]_{x^+=0}$  of two scalar observables would vanish on the grounds of microcausality principle in relativistic field theory for  $x^\perp \neq 0$  when  $x^2|_{x^+=0}$  is spacelike. Its value would hence be proportional to  $\delta^2(x^\perp)$  and a finite number of its derivatives, implying locality only in  $x^\perp$  but not necessarily so in  $x^-$ . Similar arguments in the *instant form* theory lead to the locality in all the three spatial coordinates. In view of the microcausality both of the commutators  $[A(x), B(0)]_{x^+=0}$  and  $[A(x), B(0)]_{x^0=0}$  are nonvanishing only on the light-cone,  $x^2 = 0$ . The possibility of nonlocality in the longitudinal direction in the *front form* treatment seems to allow us to display in some cases the structures parallel to those found in string theory (Sec. 4.6).

We note that in the LF quantization we time order with respect to  $\tau$  (which is a monotonic parameter as well) rather than  $t$ . The microcausality principle, however, ensures that the retarded commutators  $[A(x), B(0)]\theta(x^0)$  and  $[A(x), B(0)]\theta(x^+)$ , which appear [22] in the S-matrix elements of relativistic field theory, do not lead to disagreements in the two formulations. In the regions  $x^0 > 0, x^+ < 0$  and  $x^0 < 0, x^+ > 0$ , where the commutators seem different the  $x^2$  is spacelike and both of them vanish. Hence, admitting [23] the microcausality principle to hold, the LF hyperplane seems *equally valid and appropriate* as the conventional one of the *instant form* theory for the canonical quantization. This is demonstrated to be so, for example, in the context of SSB, SM, CSM, and QCD in covariant gauges discussed in this article.

We note that the hyper planes  $x^\pm = 0$  define the characteristic surfaces of hyperbolic partial differential equation. It is known from their mathematical theory [24] that a solution exists if we specify the (Cauchy) initial data on both of the hyperplanes. From the actual computations in the *front form* theory we come to conclusion [16] that (barring some massless theories) it is sufficient in the canonical quantization of the *front form* theory to select one of the hyperplanes. The information on the commutators on the other characteristic hyperplane seems to be already contained [15] in the quantized theory.

A distinguishing feature of the *front form* theory is that it gives rise to a constrained dynamical system [25]. After the elimination of the phase space constraints in the Hamiltonian formulation it leads to an appreciable reduction in the number of independent field operators which would describe the Hilbert space of the theory. The vacuum structure, for example, then becomes more tractable and the computation of physical quantities simpler. This is, for example, verified [15, 16, 17] in the studies of the LF quantized SM, CSM, and QCD in covariant gauges reviewed in Secs. (6, 7).

## 1.2 LF Poincaré and IMF Generators. LF Spin Operator

The structure of the *LF phase space* is different from that of the one in the conventional theory. A different description of the same physical content, compared to that found in the conventional treatment, may emerge in the *front form* theory. For example, the SSB gets a different description [32, 10] and the broken continuous symmetry is now inferred from the study of the residual unbroken symmetry of the LF Hamiltonian operator while the symmetry of the LF vacuum remains intact. However, the expression which counts the number of Goldstone bosons present in the *front form* theory, comes out to be the same as that found in the the discussion of equal-time quantized theory. A new proof of the Coleman's theorem [26] on the absence of the Goldstone bosons in two dimensional theory also emerges [32, 10]. The LF vacuum is generally found to be simpler and in many cases the interacting theory vacuum is seen to coincide with the perturbation theory one [27].

Another important advantage pointed out by Dirac of the *front form* theory is that here *seven* out of the ten Poincaré generators are *kinematical*, e.g., they leave the plane  $x^+ = 0$  invariant [1]. In the standard notation, viz.,  $K_i = -M^{0i}$ ,  $J_i = -(1/2)\epsilon_{ijk}M^{kl}$ ,  $i, j, k = 1, 2, 3$ , they are :  $P^+, P^1, P^2, M^{12} = -J_3, M^{+-} = M^{03} = -K_3, M^{1+} = (K_1 + J_2)/\sqrt{2}$ , and  $M^{+2} = (K_2 - J_1)/\sqrt{2}$ . In the conventional theory on the other hand only six such ones,  $\vec{P}$  and  $M^{ij} = -M^{ji}$ , leave the hyperplane  $x^0 = 0$  invariant. Expressed otherwise, the generator  $K_3$  is dynamical one in the *instant form* theory but it turns out to be kinematical on the LF in the sense that there it generates [15] simply the scale transformations of the LF components of  $P^\mu$  and  $M^{\mu\nu}$ , and  $x^\mu$ , with  $\mu, \nu = +, -, 1, 2$ .

There is also an interesting correspondence of the LF components of the Poincaré generators with the generators in the IMF. Consider the inertial frame  $S'$  moving along the 3-axis with velocity  $v/c = \tanh \eta$  relative to the inertial frame  $S$ . From  $(M'^{\mu\nu}, P'^\mu) = \exp(-i\eta K_3) (M^{\mu\nu}, P^\mu) \exp(i\eta K_3)$  we derive (Appendix A)

$$\begin{aligned} J'_1 &= J_1 \cosh \eta + K_2 \sinh \eta, & J'_2 &= J_2 \cosh \eta - K_1 \sinh \eta, & J'_3 &= J_3 \\ K'_1 &= K_1 \cosh \eta - J_2 \sinh \eta, & K'_2 &= K_2 \cosh \eta + J_1 \sinh \eta, & K'_3 &= K_3 \\ (P_0 + P_3)' &= e^\eta (P_0 + P_3) & (P_0 - P_3)' &= e^{-\eta} (P_0 - P_3) & P'_1 &= P_1 & P'_2 &= P_2 \end{aligned} \quad (1)$$

When  $v/c \rightarrow 1(-1)$  or  $\eta \rightarrow \infty(-\infty)$  the Lorentz transformation becomes singular. However, we may define the *renormalized* generators,  $J'_a/\cosh \eta$ ,  $K'_a/\cosh \eta$ , and  $e^{\mp\eta}(P_0 \pm P_3)'$  which have well defined limits. The generators thus obtained coincide in the limit with the LF components of the Poincaré generators. We note also that to particle at rest in  $S$  corresponds the four-momentum  $p'^\mu$  in the inertial frame  $S'$ :  $p'_\mu/(m_0 \cosh \eta)$ , which tends to a null vector.

It is also worth remarking that the  $+$  component of the Pauli-Lubanski pseudo-vector  $W^\mu$  is special in that it contains only the LF kinematical generators. We may define the *LF Spin operator* by  $\mathcal{J}_3 = -W^+/P^+$ . In the massive case the other two components of  $\vec{\mathcal{J}}$ , generating together an  $SU(2)$  algebra, are shown to be  $\mathcal{J}_a = -(\mathcal{J}_3 P^a + W^a)/\sqrt{P^\mu P_\mu}$ ,  $a = 1, 2$ , which, however, do carry in them also the LF *dynamical generators*  $P^-, M^{1-}, M^{2-}$ . The case of both the massive and massless fermions is discussed in detail in Sec. 4; the general case is considered in Appendix B.

## 2 LF quantized scalar theory

### 2.1 Covariant Phase Space Factor on the LF

Some interesting insight on the *front form* quantized field theory may already be gained by considering the Lorentz invariant phase space-*LIPS* or *Covariant phase space* [28] factor, which is found relevant in the analysis of the physical processes, introduced first in the context of the covariant version of the statistical model of Fermi [29]. On the LF the dispersion relation associated with the free massive particle is  $2p^+p^- = (p^\perp p^\perp + m^2) > 0$ . It has no square root, like in the conventional one. The *signs*, for example, of  $p^+$  and  $p^-$  are *correlated* since  $p^+p^- > 0$  [30]. The *LISP* factor in the LF coordinates is thus defined as:  $\int d^4p \theta(\pm p^+) \theta(\pm p^-) \delta(p^2 - m^2) =$

$\int d^2 p^\perp dp^+ \int dp^- \theta(\pm p^+) \theta(\pm p^-) \delta(2p^+ p^- - [m^2 + p^{\perp 2}]) = \int d^2 p^\perp dp^+ \theta(p^+) / (2p^+)$ , compared to the conventional one:  $\int d^4 p \theta(\pm p^0) \delta(p^2 - m^2) = \int d^3 \vec{p} / (2E_p)$  with  $E_p = +\sqrt{\vec{p}^2 + m^2} > 0$ . A distinguishing feature in the case of the LF is thus the appearance of  $\theta(p^+) / (2p^+)$  in the phase space factor. Such considerations are relevant, for example, in writing the Fourier transform of the fields and the discussion of chiral boson theory (Sec. 3.4).

## 2.2 LF Commutator

Consider, for example, a real massive free scalar field  $\phi(\tau, x^-, x^\perp)$ , satisfying  $(\partial_\mu^2 + m^2)\phi = 0$  where  $\partial_\mu^2 = (2\partial_+ \partial_- - \partial_\perp^2)$ . For  $p^+ > 0$ , and consequently  $p^- > 0$ , the complete set of plane wave solutions of the equation of motion are  $e^{+ip \cdot x}$  and  $e^{-ip \cdot x}$  where  $p \cdot x = p^- x^+ + p^+ x^- - p^\perp x^\perp$  and  $\tau \equiv x^+$  indicates the LF-time coordinate. The Fourier transform of  $\phi$  on the LF may clearly be written as,

$$\phi(x) = \frac{1}{\sqrt{(2\pi)^3}} \int d^2 p^\perp \frac{dp^+}{\sqrt{2p^+}} \theta(p^+) [a(p^+, p^\perp) e^{-ip \cdot x} + a^\dagger(p^+, p^\perp) e^{ip \cdot x}] \quad (2.1)$$

where we have isolated  $\sqrt{2p^+}$  only for latter convenience and  $p^\perp$  as well as  $p^+$  are to be integrated from  $-\infty$  to  $\infty$ , which is very convenient when we deal with generalized functions like  $\theta(p^+)$ . In the quantized theory  $a(p)$  and  $a^\dagger(p)$  denote the creation and annihilation operators of the quantum excitations associated with the quantized field operator  $\phi$ . They are assumed to satisfy the canonical commutation relations, with the nonvanishing one given by  $[a(k), a^\dagger(p)] = \delta(k^+ - p^+) \delta^2(k^\perp - p^\perp) \equiv \delta^3(k - p)$ . The Fock space is constructed employing these operators.

The equal-LF-time commutator of the field operator can be computed by employing its Fourier transform expression

$$\begin{aligned} [\phi(x), \phi(y)]_\tau &= \frac{1}{(2\pi)^3} \int d^2 p^\perp d^2 k^\perp \frac{dp^+ dk^+ \theta(p^+) \theta(k^+)}{\sqrt{2p^+ 2k^+}} \\ &\quad \times \left[ e^{-i(p \cdot x - k \cdot y)} - e^{i(p \cdot x - k \cdot y)} \right]_{x^+ = y^+ = \tau} \delta^3(p - k) \\ &= \frac{\delta^2(x^\perp - y^\perp)}{(2\pi)} \int \frac{dp^+}{2p^+} [\theta(p^+) + \theta(-p^+)] e^{-ip^+(x^- - y^-)} \\ &= -\frac{i}{4\pi} \epsilon(x^- - y^-) \delta^2(x^\perp - y^\perp). \end{aligned} \quad (2.2)$$

Here we have used the free particle dispersion relations for  $k^\mu$  and  $p^\mu$ , made use of the delta function in the integrand, set  $[\theta(p^+) + \theta(-p^+)] = 1$  (or rather the Cauchy principal value-CPV) in the sense of the distribution theory, and used the integral representation of the *sign* function  $\epsilon(x) = 1$  or  $-1$  according as  $x > 0$  or  $x < 0$ . The equal- $\tau$  commutator obtained here, often termed the *LF commutator*, is *nonlocal* along the longitudinal direction  $x^-$ , which as we argued before is not unexpected in the *front form* theory. It vanishes for the spacelike distances, and is nonvanishing only on the light-cone for  $x^- \neq y^-$ , when we assume  $\epsilon(0) = 0$ .

## 2.3 Length Dimensions $L_\perp$ and $L_\parallel$

It is natural and suggested also, for example, from the expression of the LF commutator to introduce [9] *two distinct units of length dimensions*,  $L_\perp$  and  $L_\parallel$  in the *front form* theory. Indicating the dimension of a quantity by  $[..]$  we write:  $[x^\perp] = L_\perp$ ,  $[x^-] = L_\parallel$ ,  $[\partial_-] = 1/L_\parallel$ . Requiring that  $p^\perp \cdot x^\perp$  be dimensionless we find  $[p^\perp] = [m] = 1/L_\perp$ , if we recall the dispersion relation. Making similar arguments we find  $[p^+] = 1/L_\parallel$ ,  $[p^-] = L_\parallel / (L_\perp)^2$ ,  $[x^+] = (L_\perp)^2 / L_\parallel$  while  $[H^{IJ}] \equiv [P^-] = L_\parallel / (L_\perp)^2$  for the LF Hamiltonian and  $[\mathcal{H}^{IJ}] = 1 / (L_\perp)^4$  for the Hamiltonian density. Similar considerations apply to the other composite operators like current densities and we remark that  $\theta(x)$  and  $\epsilon(x)$  do not carry any dimensions. The dimensional analysis is useful in finding [9] the

possible counter terms required in the renormalization of the theory. From the LF commutator (2.2) we conclude that  $[\phi] = 1/L_\perp$ , which is also found to be the case for the gauge field but for the fermionic field we have  $[\psi_+] = 1/(L_\perp\sqrt{L_\parallel})$ , where  $\psi_+$  indicates the dynamical component of the fermion field.

## 2.4 LF Hamiltonian. Dirac Procedure

The free scalar theory is described by the Lagrangian density  $\mathcal{L} = \partial_+\phi\partial_-\phi - (1/2)\partial_\perp\phi\partial_\perp\phi - m^2\phi^2/2$ . It is first order in  $\partial_+\phi$  and the canonical momenta defined as  $\pi = \partial\mathcal{L}/\partial(\partial_+\phi) = \partial_-\phi$  describes a constraint on the phase space dynamics of the *front form* scalar theory. We have here a constrained dynamical system [25]. The canonical Hamiltonian density is found to be  $\mathcal{H}_c = m^2\phi^2/2$ . There is a systematic procedure<sup>1</sup> - called the Dirac method [25]- which allows us to construct the self-consistent Hamiltonian formulation, required to canonically quantize the theory with phase space constraints. The primary constraint above is written as

$$\chi \equiv (\pi - \partial_-\phi) \approx 0 \quad (2.3)$$

where  $\approx$  stands for weak equality, meaning that it should not be employed inside the Poisson brackets, but only after they have been computed.

We define next an extended Hamiltonian density  $\mathcal{H}_e = \mathcal{H}_c + u\chi$  where  $u$  is a Lagrange multiplier field. Hamilton's equations of motion employ  $H_e \equiv \int d^2x^\perp dx^- \mathcal{H}_e$  and we require the persistency condition on the constraint, e.g.,  $\{\chi(\tau, x^\perp, x^-), H_e(\tau)\} \approx 0$ . In the simple case under study we are led to a differential equation which would determine the multiplier field  $u$ . In the gauge theory considered below new secondary constraints may arise. We now include them also in the extended Hamiltonian and repeat the procedure, until no more constraints show up. For the computational purposes we may initially start with the standard Poisson brackets at equal-LF-time  $\tau$ , with the nonvanishing one defined by<sup>2</sup>  $\{\pi(\tau, x), \phi(\tau, y)\} = -\delta^3(x - y) \equiv -\delta^2(x^\perp - y^\perp)\delta(x^- - y^-)$ .

The nature of the set of constraints found in the theory is then analyzed. A constraint is first class if it has vanishing Poisson brackets with itself, with the other constraints, and with the Hamiltonian; otherwise it is a second class one. Corresponding to a first class constraint we may be required to add in the theory some appropriate and accessible gauge-fixing external constraints [25]. In the present case there is one local constraint  $\chi \approx 0$ . From the constraint matrix

$$\{\chi(\tau, x^\perp, x^-), \chi(\tau, y^\perp, y^-)\} = -2\delta^2(x^\perp - y^\perp)\partial_-\delta(x^- - y^-). \quad (2.4)$$

we conclude that it is second class by itself, since the right hand side is nonvanishing. There is, in fact, also a first class constraint in the theory in the form of the zero-momentum-mode of  $\chi$ ; we will comment on it in Sec. (2.6).

We go over now from the Poisson to the modified Poisson brackets, called Dirac brackets, which have the property that we are allowed to set  $\chi = 0$  as a strong equality relation, valid even inside the Dirac brackets. The Hamilton's equations employ [25] now the modified brackets. We construct first the *inverse* of the constraint matrix:  $\{\chi(\tau, x^\perp, x^-), \chi(\tau, y^\perp, y^-)\}^{-1} = -\delta^2(x^\perp - y^\perp)\epsilon(x^- - y^-)/4$ . The modified bracket is then defined by

$$\{f(x), g(y)\}_D = \{f(x), g(y)\} - \int \int d^3u d^3v \{f(x), \chi(u)\} \{\chi(u), \chi(v)\}^{-1} \{\chi(v), g(y)\} \quad (2.5)$$

In view of its very construction the Dirac bracket of any dynamical variable with  $\chi$  is seen to vanish identically.

It is clear that in place of  $H_e$  we may then employ the *reduced Hamiltonian* obtained by setting  $\chi \equiv (\pi - \partial_-\phi) = 0$  in it, which would also remove the Lagrange multiplier field, while  $\pi$  becomes

<sup>1</sup>See also Secs. 5,6, and Appendix C.

<sup>2</sup>In the context of the canonical quantization we mostly deal with equal- $\tau$  brackets and commutators. We will frequently suppress  $\tau$  from writing and write occasionally  $x$  to indicate the set  $(x^\perp, x^-)$ .

now a dependent variable, e.g., removed from the theory. For the independent field  $\phi$  which survives in the *front form* scalar theory here considered we find

$$\{\phi(\tau, x), \phi(\tau, y)\}_D = -\frac{1}{4}\epsilon(x^- - y^-)\delta^2(x^\perp - y^\perp) \quad (2.6)$$

The Hamilton's equation:  $\dot{\phi}(\tau, x) = \{\phi(\tau, x), H_c\}_D$ , where an overdot indicates the derivation with respect to  $\tau$ , does recover also the Lagrange equation. The theory is canonically quantized by the correspondence  $i\{f, g\}_D \rightarrow [f, g]$ , the commutator of the corresponding quantized operators. The Hamilton's equations correspond to the equations of motion of field operators, e.g.,  $idf/d\tau = [f, H]$ , in the Heisenberg picture. The commutator of the scalar field operators on the LF is thus given by

$$[\phi(\tau, x), \phi(\tau, y)] = \frac{-i}{4}\epsilon(x^- - y^-)\delta^2(x^\perp - y^\perp) \quad (2.7)$$

which is the same as found above by the simple arguments based on the Fourier expansion of the field in the *front form* theory. Employing this commutator we recover in the present case the Lagrange equation of motion for the field operator as well.

## 2.5 Scalar Field Propagator in momentum space

The Fourier expansion (2.1) may also be regarded as furnishing the momentum space realization of the commutator (2.7) and the propagator in momentum space is easily derived. The propagator in configuration space is defined by

$$\langle 0|T(\phi(x)\phi(0))|0\rangle = \theta(\tau)\langle 0|(\phi(x)\phi(0))|0\rangle + \theta(-\tau)\langle 0|(\phi(0)\phi(x))|0\rangle. \quad (2.8)$$

It follows that

$$\begin{aligned} \langle 0|T(\phi(x)\phi(0))|0\rangle &= \frac{1}{(2\pi)^3} \int d^3p \frac{\theta(p^+)}{2p^+} [\theta(\tau)e^{-ip \cdot x} + \theta(-\tau)e^{ip \cdot x}] \\ &= \frac{i}{(2\pi)^4} \int d^3p d\lambda e^{-i(\lambda\tau + p^+x^- - p^\perp x^\perp)} \frac{[\theta(p^+) + \theta(-p^+)]}{(m^2 + p^\perp p^\perp - 2p^+\lambda - i\epsilon)} \\ &= \frac{i}{(2\pi)^4} \int d^4p \frac{e^{-ip \cdot x}}{(p^2 - m^2 + i\epsilon)} \end{aligned} \quad (2.9)$$

Here we have used the integral representations<sup>3</sup> of  $\theta(\pm\tau)$  and performed the well known standard manipulations. The factor  $[\theta(p^+) + \theta(-p^+)]$  in the integrand has been set to unity and the dummy integration variable  $\lambda$  has been renamed as  $p^-$  for convenience in the last line. The  $d^4p$  stands for  $d^2p^\perp dp^+ dp^-$ , with the understanding, as is clear from the derivation above, that the integration over the  $p^-$  has to be performed first. The range of integration is from  $-\infty$  to  $\infty$  for all of these variables.

The momentum space representations of the energy-momentum tensor are also found easily and we check that  $N(p) = a^\dagger(p)a(p)$  has the usual interpretation of a number operator. In fact,

$$\begin{aligned} H^{lf}_c &\equiv P^- = \int d^2x^\perp dx^- : \left[ \frac{m^2}{2}\phi^2 + \frac{1}{2}\partial_\perp\phi\partial_\perp\phi \right] : \\ &= \frac{1}{2} \int d^2p^\perp dp^+ \theta(p^+) : [a^\dagger(p)a(p) + a(p)a^\dagger(p)] : \frac{m^2 + p^\perp p^\perp}{2p^+} \\ &= \int d^2p^\perp dp^+ \theta(p^+) [a^\dagger(p)a(p)] p^- \\ P^+ &= \int d^2x^\perp dx^- : (\partial_- \phi)^2 : \\ &= \int d^2p^\perp dp^+ \theta(p^+) [a^\dagger(p)a(p)] p^+ \end{aligned} \quad (2.10)$$

---

<sup>3</sup> $\theta(\tau)e^{-ip^-\tau} = 1/(2i\pi) \int d\lambda e^{-i\lambda\tau}/(p^- - \lambda - i\epsilon)$

## 2.6 First class constraint. Symmetry in $x^+$ and $x^-$

It is worth making an *important remark*. There is, in fact, present [31] in the scalar theory discussed above *still another constraint which is first class*. We easily show that the zero-longitudinal-momentum mode  $\sqrt{2\pi}\tilde{\chi}(\tau, k^+ = 0) = \int dx^- \chi$ , represents a first class constraint in the theory. For example, considering for simplicity the two dimensional theory, (2.4) reads in the momentum space as

$$\{\tilde{\chi}(\tau, k^+), \tilde{\chi}(\tau, p^+)\} = -2ik^+ \delta(k^+ + p^+). \quad (2.11)$$

It clearly indicates the presence of the first class constraint  $\tilde{\chi}(\tau, k^+ = 0) \approx 0$  in the theory. Such a constraint or symmetry requires us to introduce in the theory an external (gauge-fixing) constraint [25], such that the pair becomes a second class set. We will take advantage of this gauge freedom in order to decompose the scalar field into the *bosonic condensate* variable and the quantum fluctuation field. When combined with the *standard* Dirac procedure it allows us to build [32, 10] a description of the SSB mechanism on the LF.

We also note that the *front form* formulation of *relativistic* theory is inherently symmetrical with respect to  $x^+$  and  $x^-$  and it is a matter of *convention* that we take the plus component as the LF-time while the other as a spatial coordinate. The theory quantized at  $x^+ = \text{const.}$  hyperplanes seems already to incorporate [15] in it the information on the equal- $x^-$  commutation relations. For example, we easily derive from (2.1) the following equal- $x^-$  commutator

$$[\phi(x^+, x^-, x^\perp), \phi(y^+, x^-, y^\perp)] = \frac{1}{(2\pi)^3} \int d^2 p^\perp \frac{dp^+ \theta(p^+)}{2p^+} \left[ e^{-ip^-(x^+ - y^+) + ip^\perp(x^\perp - y^\perp)} - e^{ip^-(x^+ - y^+) - ip^\perp(x^\perp - y^\perp)} \right]. \quad (2.12)$$

In view of the free particle dispersion relation we may replace the measure  $dp^+ \theta(p^+)/2p^+$  by  $dp^- \theta(p^-)/2p^-$ . The equal- $x^-$  commutator is then given by  $(-i/(4\pi))\epsilon(x^+ - y^+)\delta^2(x^\perp - y^\perp)$ . In two dimensional space-time it is customary to define the right and the left movers by  $\phi(0, x^-) \equiv \phi^R(x^-)$ , and  $\phi(x^+, 0) \equiv \phi^L(x^+)$ . We find  $[\phi^R(x^-), \phi^R(y^-)] = (-i/4)\epsilon(x^- - y^-)$  while  $[\phi^L(x^+), \phi^L(y^+)] = (-i/4)\epsilon(x^+ - y^+)$ . The symmetry under discussion seems responsible for appreciable simplifications in the *front form* quantized theory.

## 3 SSB Mechanism, Topological Kink Solution, and Chiral Boson theory on LF

### 3.1 SSB in two dimensional scalar theory

The conventional *instant form* description of the tree level SSB is based on the space and time independent solutions of the Lagrange equation,  $\phi_{\text{class}} \equiv \omega$ , such that they also minimize the Hamiltonian functional; based on the (external) physical considerations. We do not apparently have much physical intuition on the LF to avail of such arguments. The constrained dynamical system on the LF seems, however, to already contain in it the corresponding relevant constraints. For simplicity we consider first the two dimensional theory<sup>4</sup> with  $\mathcal{L} = (\partial_+ \phi)(\partial_- \phi) - V(\phi)$

This is probably the simplest example of a constrained dynamical system in the context of field theory. It is reasonable to expect that the well tested Dirac procedure, when applied to it, *must* result in a satisfactory description of SSB on the LF.

The Lagrange equation,  $2\dot{\phi}' = -V'(\phi)$ , is of first order in LF-time  $\tau$ . The left hand side remains unaltered under  $\phi \rightarrow \phi + c(\tau)$  and  $\phi = \text{const.}$  are clearly possible solutions<sup>5</sup>. Integrating over the space variable and assuming appropriate boundary conditions we are led to the following

<sup>4</sup>Here  $\tau = (x^0 + x^1)/\sqrt{2}$ ,  $x \equiv x^- = (x^0 - x^1)/\sqrt{2}$ . An overdot indicates the LF-time derivative while a prime indicates derivative with respect to  $x^-$ ; the generalization to 3 + 1 dimensions is discussed in Sec. (3.2).

<sup>5</sup>The self-dual *kink* solution which depends on  $x^-$  as well is discussed in Sec. (3.3).



constraint [32, 10] on the potential

$$\int dx^- \frac{\delta V(\phi)}{\delta \phi} = 0. \quad (3.1)$$

We show now that this constraint is also present on the phase space and in the quantized theory. The description of SSB then follows from the discussion on the structure of the Hilbert space.

In order to take care of the first class constraint  $\int dx^- \chi \approx 0$  mentioned in Sec. (2.6) we make the following *separation* of the dynamical (collective) bosonic *condensate* variable  $\omega(\tau)$  from the (quantum) fluctuation variable  $\varphi(\tau, x)$

$$\phi(\tau, x) = \omega(\tau) + \varphi(\tau, x). \quad (3.2)$$

Here we also set  $\int dx^- \varphi(\tau, x) = 0$  so that the fluctuation field carries no zero-longitudinal-momentum mode in it. The *separation thus corresponds in a sense to an external gauge-fixing constraint* which we must impose [25] in the theory. It was introduced [32] originally on physical considerations and  $\omega$  was termed as the dynamical *bosonic condensate* variable.

We apply now the standard Dirac procedure to construct LF Hamiltonian formulation. The canonically quantized theory results [10, 32] in the following commutators

$$[\varphi(x, \tau), \varphi(y, \tau)] = -\frac{i}{4}\epsilon(x - y), \quad (3.3)$$

$$[\omega(\tau), \varphi(x, \tau)] = 0. \quad (3.4)$$

and for  $V(\phi) = (\lambda/4)(\phi^2 - m^2/\lambda)^2$ , with a negative sign for the mass term and  $\lambda \geq 0$ ,  $m \neq 0$ , the LF Hamiltonian is given by

$$H^{lf} \equiv P^- = \int dx \left[ \omega(\lambda\omega^2 - m^2)\varphi + \frac{1}{2}(3\lambda\omega^2 - m^2)\varphi^2 + \lambda\omega\varphi^3 + \frac{\lambda}{4}\varphi^4 \right]. \quad (3.5)$$

We recover also the *constraint equation* (3.1) now as a second class constraint on the phase space:

$$\omega(\lambda\omega^2 - m^2) + \frac{1}{R} \int_{-R/2}^{R/2} dx \left[ (3\lambda\omega^2 - m^2)\varphi + \lambda(3\omega\varphi^2 + \varphi^3) \right] = 0 \quad (3.6)$$

where  $R \rightarrow \infty$  and the Cauchy principle value of  $\int_{-\infty}^{\infty} dx f(x)$  is defined by  $\lim_{R \rightarrow \infty} \int_{-R/2}^{R/2} dx f(x)$ .

The commutation relations indicate that the operator  $\omega$  is a c-number or a background field. Eliminating  $\omega$  would lead to *LF Hamiltonian which is nonlocal* [32, 10] along the longitudinal coordinate  $x^-$  even though the scalar theory written in the conventional coordinates is local.

At the tree or classical level,  $\varphi$  are bounded ordinary functions in  $x^-$  and when  $R \rightarrow \infty$  only the first term survives in the constraint equation leading to  $\omega(\lambda\omega^2 - m^2) = 0$ . This result is the same as that obtained in the conventional theory. There, however, it is essentially added to the theory, on physical considerations, which require the energy functional to attain its minimum (extremum) value. The *stability property*, say, of a particular constant solution may be inferred as usual from the analysis of the classical partial differential equation of motion. For example,  $\omega = 0$  is shown to be an unstable solution on the LF for the potential considered above, while the other two solutions with  $\omega \neq 0$  give rise to the stable phases. A similar analysis, it is clear, of the corresponding *partial* differential equations in the conventional coordinates can also be made; the Fourier transform theory is convenient to use. Also the new ingredient in the form of the constraint equation on the LF does have its counterpart in the conventional *instant form* framework as is shown in [18]. It is *remarkable* that the *front form* theory seems to contain inside it all the necessary ingredients in order to describe the SSB, when we follow the Dirac procedure to handle the constrained LF dynamics of the scalar field.

We could have employed the DLCQ [33], including the condensate term also in it. The existence of the *continuum* limit of DLCQ theory adding to it also the dynamical condensate variable was

demonstrated [32, 34], contradicting the then prevalent notion on the contrary<sup>6</sup>. The demonstration assures [35] us of the self-consistency of the *front form* theory itself. In the infinite volume limit [10, 34], we do obtain the same results. However, in the theory described in finite volume, the commutator of  $\omega$  with  $\varphi$  is found nonvanishing and as such it is an *operator*; only when  $R \rightarrow \infty$  does it become a classical background field.

It is worth stressing that in our discussion the condensate variable is introduced as a dynamical variable. The Dirac *procedure must decide* whether it comes out to be c- or q-number. In the discussions of the bosonized SM and CSM models the operator  $\omega$  is not a background field, like in the scalar theory. It turns out to be an operator and plays an important role in describing the structure of the Hilbert space and the degenerate vacua in these gauge theories (Sec. 6).

The field commutator obtained above can be realized in momentum space through the Fourier transform of  $\varphi$ :  $\varphi(x, \tau) = (1/\sqrt{2\pi}) \int dk \theta(k) [a(k, \tau) e^{-ikx} + a^\dagger(k, \tau) e^{ikx}]/(\sqrt{2k})$ , where  $k^+ \equiv k$  and the operators  $a(k, \tau)$  and  $a^\dagger(k, \tau)$  satisfy the canonical equal- $\tau$  commutation relations, with the nonvanishing one given by  $[a(k, \tau), a^\dagger(k', \tau)] = \delta(k - k')$ .

The (perturbative) vacuum state is defined by  $a(k, \tau)|vac\rangle = 0$ ,  $k > 0$ . The tree level description of the SSB is given as follows. The values of  $\omega = \langle |\phi| \rangle_{vac}$  obtained from  $V'(\omega) = 0$  characterize the different vacua in the theory. Distinct Fock spaces corresponding to different values of  $\omega$  are built as usual by applying the creation operators on the corresponding vacuum state. The  $\omega = 0$  corresponds to *symmetric phase* since the Hamiltonian operator is then symmetric under  $\varphi \rightarrow -\varphi$ . For  $\omega \neq 0$  this symmetry is violated and the system is said to be in a *broken or asymmetric phase*.

The constraint equation (3.6) also shows that the value of  $\omega$  would be altered from its tree level value in view of the quantum corrections, arising from the other terms. The renormalization of the two dimensional scalar theory was discussed [18] to one-loop order by employing the Dyson-Wick expansion based on LF-time ordering. It was found that it is convenient to derive [18] the renormalized constraint equation instead of solving the constraint equation first, which would require the difficult job of dealing with nonlocal and nonlinear Hamiltonian.

In the supernormalizable theory here the two renormalized equations, viz, the mass renormalization condition and the renormalized constraint equation, allow us to study [18] the phase transition in the two dimensional scalar theory, which was conjectured long time ago by Simon and Griffiths [36].

### 3.2 Spontaneously broken continuous symmetry

The *extension to 3+1 dimensions* and to the global *continuous symmetry* is straightforward [10]. Consider real scalar fields  $\phi_a (a = 1, 2, \dots, N)$  which form an isovector of global internal symmetry group  $O(N)$ . We now write  $\phi_a(x, x^\perp, \tau) = \omega_a + \varphi_a(x, x^\perp, \tau)$  and the Lagrangian density is  $\mathcal{L} = [\dot{\varphi}_a \varphi'_a - (1/2)(\partial_\perp \varphi_a)(\partial_\perp \varphi_a) - V(\phi)]$ . The Taylor series expansion of the constraint equations  $\beta_a = 0$  gives a set of coupled equations  $RV'_a(\omega) + V''_{ab}(\omega) \int dx \varphi_b + V'''_{abc}(\omega) \int dx \varphi_b \varphi_c / 2 + \dots = 0$ . Its discussion at the tree level leads to the conventional theory results. The LF symmetry generators are found to be  $G_\alpha(\tau) = -i \int d^2 x^\perp dx \varphi'_c (t_\alpha)_{cd} \varphi_d = \int d^2 k^\perp dk \theta(k) a_c(k, k^\perp)^\dagger (t_\alpha)_{cd} a_d(k, k^\perp)$  where  $\alpha, \beta = 1, 2, \dots, N(N-1)/2$ , are the group indices,  $t_\alpha$  are hermitian and antisymmetric generators of  $O(N)$ , and  $a_c(k, k^\perp)^\dagger$  ( $a_c(k, k^\perp)$ ) is creation (destruction) operator, contained in the momentum space expansion of  $\varphi_c$ . These are to be contrasted with the generators in the equal-time theory,  $Q_\alpha(x^0) = \int d^3 x J^0 = -i \int d^3 x (\partial_0 \varphi_a) (t_\alpha)_{ab} \varphi_b - i (t_\alpha \omega)_a \int d^3 x (d\varphi_a/dx_0)$ . All the symmetry generators thus annihilate the LF vacuum and the SSB is now seen in the broken symmetry of the quantized theory Hamiltonian. The expression which counts the number of Goldstone bosons in the *front form* theory is found to be identical to that in the conventional theory. In contrast, the first term on the right hand side of  $Q_\alpha(x^0)$ , which is similar to the one on the LF, does annihilate the conventional theory vacuum but the second term gives now non-vanishing contributions for some of the (broken) generators. The symmetry of the conventional theory vacuum is thereby broken while the quantum Hamiltonian remains invariant. The *physical content* of SSB in the

<sup>6</sup>See, T. Maskawa and K. Yamawaki, Prog. Theo. Phys. 56 (1976) 270; K. Nakanishi and K. Yamawaki, Nucl. Phys. B122 (1977) 15. A history of the so called zero mode problem is traced [10] in hep-th/9312064.

*instant form* and the *front form*, however, is the same though achieved by different descriptions. Alternative proof on the LF, in two dimensions, can be given of the Coleman's theorem related to the absence of Goldstone bosons; we are unable [10] to implement the second class constraints over the phase space. The tree level Higgs mechanism may also be discussed straightforwardly [10]. We remark that the simplicity of the LF vacuum is in a sense compensated by the involved nonlocal Hamiltonian. The latter, however, may be treatable using advance computational techniques. Also in connection with renormalization it may not be necessary [18]; we may instead obtain the renormalized constraint equations.

### 3.3 Kink solution and Topological quantum number

The classical Lagrange equation of the two dimensional self-interacting theory,  $2\partial_-\partial_+\phi = -V'(\phi)$ , with the  $V(\phi)$  given above, is known to have finite energy topological soliton solutions [37] called *kink* solutions. The theory has an internal symmetry,  $\phi \rightarrow -\phi$ . They can be recovered in the *front form* theory as well. The *kink* corresponds to the *self-dual* solution satisfying  $\partial_-\phi = -\partial_+\phi$  and given by

$$\phi_{kink} = \pm \frac{m}{\sqrt{\lambda}} \tanh \left[ \frac{m}{2} (x^+ - x^-) \right] \quad (3.7)$$

where the upper (lower) sign corresponds to the kink (anti-kink) solutions. The kink on the LF carries both the LF energy and longitudinal momentum such that  $P^+ = P^-$  and<sup>7</sup> its mass is determined to be  $M = \sqrt{2P^+P^-} = \sqrt{8}m^3/(3\lambda)$ . The kink interpolates between the two vacua of the theory:  $\phi_{kink}(0, x^- = \infty) = -m/\sqrt{\lambda}$  and  $\phi_{kink}(0, x^- = -\infty) = m/\sqrt{\lambda}$ . The topological charge may be defined by  $Q = \int dx^- j^+$  where  $j^\mu = -(\sqrt{\lambda}/(2m))\epsilon^{\mu\nu}\partial_\nu\phi$ , with  $\epsilon^{+-} = -\epsilon^{-+} = \epsilon_{01} = 1$ , is the conserved topological current density. The topological charges of kink, anti-kink and vacuum solutions are 1, -1, and 0 respectively. The  $Q$  is absolutely conserved prohibiting the decay of the kink into vacuum. Similar (topological) quantum numbers on the LF arise also, for example, in the context of the structure of the degenerate vacua in the canonical quantization of SM and CSM models discussed below.

### 3.4 Chiral Boson theory on the LF

The chiral boson (or self-dual scalar) field in 1+1 dimensions plays an important role, for example, in the formulation of string theories [38], in the description [39] of boundary excitations of the quantum Hall state, and in a number of two-dimensional statistical systems which are related to the Coulomb-gas model.

We recall that the free massive theory with  $\mathcal{L} = \partial^\mu\phi\partial_\mu\phi/2 - m^2\phi^2/2$  has the LF Hamiltonian  $m^2\phi^2/2$ . The dispersion relation  $2p^+p^- = m^2 > 0$  governs the *correlation* between the signs of  $p^+$  and  $p^-$ . It ceases to exist in the massless theory where, at the classical level, a chiral boson solution,  $\partial_0\phi = \partial_1\phi$  (and an anti-chiral one,  $\partial_0\phi = -\partial_1\phi$ ) is obtained. Several quantized theory models [40, 41, 42] of chiral boson have been proposed. The *front form* theory of chiral boson looks more appropriate and transparent [43] when compared to the conventional one.

The Floreanini and Jackiw (FJ) model [41] is based on the following *manifestly non-covariant Lagrangian*

$$\begin{aligned} \mathcal{L} &= (\partial_0\phi - \partial_1\phi)\partial_1\phi \\ &= \frac{1}{2}\eta^{\mu\nu}\partial_\mu\phi\partial_\nu\phi - \frac{1}{2}(\partial_0\phi - \partial_1\phi)^2. \end{aligned} \quad (3.8)$$

where  $\phi$  is a real scalar field and  $\eta^{00} = -\eta^{11} = 1$ ,  $\eta^{01} = \eta^{10} = 0$ .

In the *instant form* frame work it leads [41, 44] to the following equal-time commutator

$$[\phi(x^0, x^1), \phi(x^0, y^1)] = \frac{-i}{4}\epsilon(x^1 - y^1). \quad (3.9)$$

---

<sup>7</sup> $P^- = \int dx^- V(\phi)$  and  $P^+ = \int dx^- (\partial_-\phi)^2$ .

The commutator is nonvanishing, is nonlocal, and violates the microcausality principle, contrary to what we encounter usually in the conventional theory [23]. These *objections disappear* when we consider the theory quantized in the LF coordinates.

We will consider a *modified* FJ chiral boson model with the following Lagrangian density written in the LF coordinates

$$\begin{aligned}\mathcal{L} &= (\partial_+\phi - \frac{1}{\alpha}\partial_-\phi) \partial_-\phi \\ &= \frac{1}{2}\eta^{\mu\nu}\partial_\mu\phi\partial_\nu\phi - \frac{1}{\alpha}(\partial_-\phi)^2,\end{aligned}\quad (3.10)$$

where  $\eta^{+-} = \eta^{-+} = 1, \eta^{++} = \eta^{--} = 0$  and  $\alpha$  is a fixed parameter. For  $\alpha = 1$  it coincides with (3.8) in the conventional coordinates.

The LF quantization of the scalar theory with a potential term included in it has been discussed in Sec. (2.4). From (3.10) we derive

$$\begin{aligned}H^{lf} &= \int dx^- \frac{1}{\alpha}(\partial_-\phi)^2 \\ [\phi(\tau, x^-), \phi(\tau, y^-)] &= \frac{-i}{4}\epsilon(x^- - y^-)\end{aligned}\quad (3.11)$$

The LF commutator (3.11), which is nonlocal in  $x^-$  and nonvanishing only on the light-cone, does not conflict with the microcausality (Sec. 1.1 and [23]) unlike (3.9).

The Heisenberg equation of motion for the field operator is

$$\partial_+\phi = \frac{1}{i} [\phi, H^{lf}] = \frac{1}{\alpha}\partial_-\phi \quad (3.12)$$

and the Lagrange equation

$$\partial_- \left[ \partial_+\phi - \frac{1}{\alpha}\partial_-\phi \right] = 0. \quad (3.13)$$

is recovered.

The commutator (3.11) can be realized in momentum space through the following Fourier transform ( $x^+ \equiv \tau$ )

$$\phi(x^+, x^-) = \frac{1}{\sqrt{2\pi}} \int dk^+ \frac{\theta(k^+)}{\sqrt{2k^+}} \left[ a(x^+, k^+) e^{-ik^+x^-} + a^\dagger(x^+, k^+) e^{ik^+x^-} \right], \quad (3.14)$$

if the operators  $a$  and  $a^\dagger$  are assumed to satisfy the equal- $\tau$  canonical commutation relations, with the nonvanishing one given by  $[a(x^+, k^+), a^\dagger(x^+, p^+)] = \delta(k^+ - p^+)$ . On using the equation of motion (3.12) we derive easily

$$a(x^+, k^+) = e^{-ik^-x^+} a(k^+), \quad a^\dagger(x^+, k^+) = e^{ik^-x^+} a^\dagger(k^+). \quad (3.15)$$

where we set

$$k^- = \frac{1}{\alpha}k^+, \quad \text{implying} \quad 2k^+k^- = \frac{2}{\alpha}(k^+)^2. \quad (3.16)$$

The dispersion relation for the free FJ chiral boson is different from that for a free scalar particle with (finite  $k^+$  but) vanishing mass, except for when  $|\alpha| \rightarrow \infty$ .

The Fourier transform now assumes the form

$$\phi(x^+, x^-) = \frac{1}{\sqrt{2\pi}} \int dk^+ \frac{\theta(k^+)}{\sqrt{2k^+}} \left[ a(k^+) e^{-ik \cdot x} + a^\dagger(k^+) e^{ik \cdot x} \right]. \quad (3.17)$$

where  $k \cdot x \equiv k^-x^+ + k^+x^- = k^+(x^- + x^+)/\alpha$  and the nonvanishing commutator satisfies  $[a(k^+), a^\dagger(p^+)] = \delta(k^+ - p^+)$ .

The components of the classical canonical energy-momentum tensor  $T^{\mu\nu}$  following from the noncovariant Lagrangian density (3.10) are found to be

$$\begin{aligned} T^{+-} = -T^{-+} &= \frac{1}{\alpha} T^{++} = \frac{1}{\alpha} (\partial_- \phi)^2, \\ T^{--} &= (\partial_+ \phi)^2 - \frac{2}{\alpha} (\partial_+ \phi)(\partial_- \phi). \end{aligned} \quad (3.18)$$

The on shell conservation equations

$$\partial_\mu T^{\mu\pm} = 2(\partial_\mp \phi) \partial_- \left[ \partial_+ \phi - \frac{1}{\alpha} \partial_- \phi \right] = 0 \quad (3.19)$$

may be easily checked. They allow us to define, if the surface integrals can be dropped, the conserved translation generators  $P^\pm$

$$P^+ = \int dx^- : T^{++} : = \int dx^- : (\partial_- \phi)^2 : = \int dk^+ \theta(k^+) N(k^+) (k^+) \quad (3.20)$$

and

$$P^- \equiv H^{lf} = \int dx^- : T^{+-} : = \frac{1}{\alpha} P^+ \quad (3.21)$$

where  $N(k^+) = a^\dagger(k^+)a(k^+)$  is the number operator and  $:$  indicates the normal ordering.

From (3.18) and in virtue of  $(T^{+-} + T^{-+}) = 0$  we may derive the following relation

$$\partial_+ [x^- T^{++} + x^+ T^{+-}] + \partial_- [x^- T^{-+} + x^+ T^{--}] = 0. \quad (3.22)$$

which is valid on shell. We may hence define another conserved generator

$$M = x^+ P^- + \int dx^- x^- T^{++} \quad (3.23)$$

The generators  $M, P^+, P^-$  form a closed algebra:  $[M, P^+] = -iP^+$ ,  $[M, P^-] = -iP^-$ , and  $[P^+, P^-] = 0$ . The operator  $M$  thus generates the scale (boost) transformations on  $P^\pm$  by the same amount which leaves  $P^+/P^-$  invariant. The mass operator  $2P^+P^-$ , however, gets scaled and is not invariant under  $M$ . The usual (kinematical) Lorentz boost generator  $M^{+-} \equiv -x^+ P^- + \int dx^- x^- T^{++}$  has similar properties. It is, however, as seen from (3.19), is *not conserved* in the manifestly *noncovariant model* under consideration. The Lagrange equation is shown to be form invariant under the infinitesimal transformation [41, 45]  $\phi \rightarrow \phi + \epsilon(x^- + x^+/\alpha)\partial_- \phi$  generated by  $M$ .

In the limit when  $|\alpha| \rightarrow \infty$  we find  $\phi \rightarrow \phi_R(x^-)$  while  $H^{lf} \rightarrow 0$ , which corresponds to the LF Hamiltonian of free massless scalar theory, Sec. (2.6). The field  $\phi_R$  satisfies:  $[\phi_R(x^-), \phi_R(y^-)] = -i\epsilon(x^- - y^-)/4$ . The limiting case is thus seen to describe a right (moving) chiral boson theory with the Lagrangian density as given in (3.10).

An alternative form of the Lagrangian density may also be employed in our context. We recall that in the quantization of gauge theory it is found useful (Sec. 5) to introduce an auxiliary field  $B(x)$  of canonical mass dimension two (in 3+1 dimensions) and add  $(B\partial_\mu A^\mu + \alpha B^2)$  as the gauge-fixing term to the Lagrangian density. In the two dimensional theory under consideration it is also possible to follow this procedure, since the corresponding  $B(x)$  field here carries the canonical mass dimension one. The discussion parallel to the one given above may thus be based also on the following [46, 47] Lagrangian density

$$\mathcal{L} = \frac{1}{2} \eta^{\mu\nu} \partial_\mu \phi \partial_\nu \phi + \sqrt{2} B(x) (\partial_- \phi) + \frac{\alpha}{2} B(x)^2. \quad (3.24)$$

The elimination of the auxiliary field using its equation of motion leads to (3.10) and the conclusions reached are the same.

We make only brief comments on other models. Siegel's [40] theory which employs

$$\mathcal{L} = \frac{1}{2}\eta^{\mu\nu}\partial_\mu\phi\partial_\nu\phi + B(x)(\partial_0\phi - \partial_1\phi)^2 \quad (3.25)$$

is afflicted by anomaly which is to be eliminated by the addition of a Wess-Zumino term. The resulting theory does not describe [48] pure chiral bosons since they are coupled to the gravity. In this model the auxiliary field carries vanishing canonical dimension and, for example, a  $B^2$  term cannot be added without introducing the dimensionful parameters.

The model based on the idea of implementing the chiral constraint through a linear constraint [42, 46],

$$\mathcal{L} = \frac{1}{2}\eta^{\mu\nu}\partial_\mu\phi\partial_\nu\phi + B_\mu(\eta^{\mu\nu} - \epsilon^{\mu\nu})\partial_\nu\phi, \quad (3.26)$$

where  $B_\mu$  is Lagrange multiplier field, does not seem to exhibit physical excitations [49]. We note that the field  $B_\mu$  carries dimension one and that this is the usual procedure in the classical theory which, however, seems to break down at the quantum level.

To summarize, the simple procedure of separating first the *condensate variable*, which in fact corresponds to a gauge-fixing condition needed on the phase space in the context of Dirac procedure, before applying the standard procedure itself, is found to be successful in describing [32, 10] the SSB, the phase transition in two dimensional scalar theory, the SSB of continuous symmetry in 3+1 dimensional theory, in furnishing a new proof of the Coleman's theorem, and in the description of (the tree level) Higgs mechanism. It is also found successful in showing [15, 16] transparently and economically the vacuum structures in the SM and CSM models as will be reviewed in Sec. 5. The self-duality constraint in the interacting theory leads to the well known *kink* solution in the *front form* theory as well. The chiral boson theory discussion becomes transparent and the LF commutator does not conflicts with the microcausality. A transparent discussion of the chiral boson theory emerges in the context of the *modified* FJ model.

We will next review the essentials of the LF quantization of the Dirac and Maxwell fields.

## 4 LF quantized Dirac field

### 4.1 Anticommutators

On the LF there is a natural decomposition of the spinor space. The LF components [50]  $\gamma^\pm$ , where  $\gamma^\pm = (\gamma^0 \pm \gamma^3)/\sqrt{2}$  have the properties  $(\gamma^\pm)^2 = (\gamma^\mp)^2 = 0$ ,  $\gamma^0\gamma^+ = \gamma^-\gamma^0$ ,  $\gamma^{+\dagger} = \gamma^-$ , and  $\gamma^+\gamma^- + \gamma^-\gamma^+ = 2I$ . We may thus introduce the hermitian projection operators  $\Lambda^\pm$

$$\Lambda^\pm = \frac{1}{2}\gamma^\mp\gamma^\pm = \frac{1}{\sqrt{2}}\gamma^0\gamma^\pm, \quad (\Lambda^\pm)^2 = \Lambda^\pm, \quad \Lambda^+\Lambda^- = \Lambda^-\Lambda^+ = 0, \quad \gamma^0\Lambda^+ = \Lambda^-\gamma^0 \quad (4.1)$$

The corresponding  $\pm$  projections of the LF Dirac spinor are  $\psi_\pm = \Lambda^\pm\psi$  and  $\bar{\psi} = \psi^\dagger\gamma^0 = \bar{\psi}_+ + \bar{\psi}_-$ ,  $\gamma^\pm\psi_\mp = 0$ ,  $\Lambda^\pm\psi_\pm = \psi_\pm$  etc. The matrix  $\Sigma_3 = \Sigma_3^\dagger = i\gamma^1\gamma^2$ ,  $\Sigma_3^2 = I$ , which commutes with  $\Lambda^\pm$  plays an important role on the LF and we note :  $(\Lambda^+ + \Lambda^-) = I$ ,  $(\Lambda^+ - \Lambda^-) = \Sigma_3\gamma_5$ ,  $\gamma_5\psi = \Sigma_3(\psi_+ - \psi_-)$ , and  $\Sigma_3\gamma^\perp\Sigma_3 = -\gamma^\perp$ .

The action of the free Dirac field is [51]

$$\begin{aligned} S &= \int d^2x^\perp dx^- \mathcal{L} \quad \text{where} \\ \mathcal{L} &= \bar{\psi}(i\{\gamma^+\partial_+ + \gamma^-\partial_- + \gamma^\perp\partial_\perp\} - m)\psi \\ &= i\sqrt{2}\psi_+^\dagger\partial_+\psi_+ + i\sqrt{2}\psi_-^\dagger\partial_-\psi_- \\ &\quad - \psi_-^\dagger(m + i\gamma^\perp\partial_\perp)\gamma^0\psi_+ - \psi_+^\dagger(m + i\gamma^\perp\partial_\perp)\gamma^0\psi_- \end{aligned} \quad (4.2)$$

It shows that only the component  $\psi_+$  carries kinetic term and the  $\psi_-$  component is nondynamical. The variation of the action with respect to  $\psi_-^\dagger$  and  $\psi_-$  leads to the constraint equation

$$2i\partial_-\psi_- = (m + i\gamma^\perp\partial_\perp)\gamma^+\psi_+ \quad (4.3)$$

and its conjugate, while for the dynamical component  $\psi_+$  we obtain the equation of motion

$$4\partial_+\psi_+ = -(m + i\gamma^\perp\partial_\perp)\gamma^-\frac{1}{\partial_-}(m + i\gamma^\perp\partial_\perp)\gamma^+\psi_+, \quad (4.4)$$

after eliminating the dependent component  $\psi_-$ . Its right hand side may be simplified to  $2(-m^2 + \partial^\perp\partial^\perp)(1/\partial_-)\psi_+$ . The canonical Hamiltonian density is easily seen to be  $\mathcal{H}_c^{lf} = \psi_+^\dagger(m + i\gamma^\perp\partial_\perp)\gamma^0\psi_-$  with  $\psi_-$  being a dependent field given by the constraint equation above. It is straightforward to verify that the equation of motion for the dynamical component  $\psi_+$  in the quantized theory is recovered as an Heisenberg equation of motion if we postulate the following anticommutation relations, which are *local in all the spatial cooriantes*.

$$\begin{aligned} \{\psi_+(\tau, x^-, x^\perp), \psi_+^\dagger(\tau, y^-, y^\perp)\} &= \frac{1}{\sqrt{2}}\Lambda^+\delta(x^- - y^-)\delta^2(x^\perp - y^\perp), \\ \{\psi_+(\tau, x^-, x^\perp), \psi_+(\tau, y^-, y^\perp)\} &= 0, \quad \{\psi_+^\dagger(\tau, x^-, x^\perp), \psi_+^\dagger(\tau, y^-, y^\perp)\} = 0. \end{aligned} \quad (4.5)$$

The same result is also derived if we follow the straightforward Dirac procedure as in the case of the scalar theory. No first class constraint, however, arises in the present case. The scale dimension of  $\psi_+$  is clearly  $[\psi_+] = 1/(L_\perp\sqrt{L_\parallel})$ . It follows from (4.3) that

$$\{\psi_-(\tau, x^-, x^\perp), \psi_+^\dagger(\tau, y^-, y^\perp)\} = \frac{1}{i4\sqrt{2}}(m + i\gamma^\perp\partial_\perp)\gamma^+\epsilon(x^- - y^-)\delta^2(x^\perp - y^\perp) \quad (4.6)$$

## 4.2 LF Spinor in momentum space and its properties

In order to write the Fourier transform we look for the complete set of linearly independent plane wave solutions of the free Dirac equation in the *front form* theory. For the massive field the signs of  $p^+$  and  $p^-$  are correlated. Choosing, say,  $p^+ > 0$  the independent set of the plane wave solutions are  $u(p)e^{-ip\cdot x}$  and  $v(p)e^{ip\cdot x}$  where the four-spinors  $u(p)$  and  $v(p)$  satisfy:  $(m - \gamma^\mu p_\mu)u(p) = 0$  and  $(m + \gamma^\mu p_\mu)v(p) = 0$ . We will make the phase convention such that  $v(p) = C\gamma^{0T}u(p)^*$ , the charge conjugate of  $u(p)$ .

A very useful form [15, 10] of the free *LF four-spinor* is given by

$$u^{(r)}(p) = N(p) \left[ \sqrt{2}p^+\Lambda^+ + (m + \gamma^\perp p_\perp)\Lambda^- \right] \tilde{u}^{(r)}, \quad (4.7)$$

where the normalization is chosen as  $N(p) = 1/(\sqrt{2}p^+m)^{1/2}$ , with  $m > 0$  and  $p^+ > 0$ . The constant spinors  $\tilde{u}^{(r)}$ , which are also the spinors in the rest frame  $\tilde{p} = (m/\sqrt{2}, m/\sqrt{2}, 0^\perp)$ , satisfy  $\gamma^0\tilde{u}^{(r)} = \tilde{u}^{(r)}$ ,  $\Sigma_3\tilde{u}^{(r)} = r\tilde{u}^{(r)}$  with  $r = \pm$ . The charge conjugate rest frame spinors satisfy  $\gamma^0\tilde{v}^r = -\tilde{v}^r$  and  $\Sigma_3\tilde{v}^{(r)} = -r\tilde{v}^{(r)}$  while

$$v^{(r)}(p) = N(p) \left[ \sqrt{2}p^+\Lambda^+ + (m - \gamma^\perp p_\perp)\Lambda^- \right] \tilde{v}^{(r)}. \quad (4.8)$$

We note that  $\gamma_5 u^{(r)}(p; m) = r u^{(r)}(p; -m)$  and  $\gamma_5 v^{(r)}(p; m) = -r v^{(r)}(p; -m)$  indicating the mass reversal property of  $\gamma_5$  upto a phase factor. Also  $\Sigma_3 u^{(r)}(p; m) = r u^{(r)}(p^+, -p^\perp; m)$  and  $\Sigma_3 v^{(r)}(p; m) = -r v^{(r)}(p^+, -p^\perp; m)$ . We do *not* introduce two spinors and work only with four-spinors and do not also employ any explicit matrix representation.

We recall that the LF *Spin operator* for the massive as well as massless particles is defined (Appendix B) by  $\mathcal{J}_3 \equiv -W^+/P^+$  where  $W^\mu$  is the Pauli-Lubanski four-vector. It contains solely the LF kinematical generators and the following useful identity can be demonstrated [15, 10]

$$\mathcal{J}_3(p) = e^{(-\frac{i}{p^+})(B_1 p^1 + B_2 p^2)} J_3 e^{(\frac{i}{p^+})(B_1 p^1 + B_2 p^2)} = J_3 - \frac{1}{p^+} (p^1 B_2 - p^2 B_1) \quad (4.9)$$

where  $\sqrt{2}B_1 = (K_1 + J_2)$  and  $\sqrt{2}B_2 = (K_2 - J_1)$  are the kinematical boost operators on the LF in the standard notation. Applying it to the spin 1/2 case<sup>8</sup> we derive ( $J_3 = \Sigma_3/2$ )

$$\begin{aligned}\mathcal{J}_3(p) &= \frac{1}{2} \left[ I + \frac{(\gamma^\perp p_\perp) \gamma^+}{p^+} \right] \Sigma_3 \\ &= J_3 + \frac{(\gamma^\perp p_\perp)}{2p^+} \gamma^+ \gamma_5 \\ \mathcal{J}_3(p) u^{(r)}(p) &= (r/2) u^{(r)}(p) \\ \mathcal{J}_3(p) v^{(r)}(p) &= -(r/2) v^{(r)}(p)\end{aligned}\quad (4.10)$$

where  $r/2 = \pm(1/2)$  are the projections of  $\vec{\mathcal{J}}(p)$  on the 3-axis in the rest frame and we used  $i(\gamma^2 p^1 - \gamma^1 p^2) = (\gamma^\perp p_\perp) \Sigma_3$ . The four-spinors are shown to satisfy the following orthogonality relations:

$$\bar{u}^{(r)}(p) u^{(s)}(p) = \delta_{rs}, \quad \bar{v}^{(r)}(p) v^{(s)}(p) = -\delta_{rs}, \quad \bar{u}^{(r)}(p) v^{(s)}(p) = 0. \quad (4.11)$$

and the following completeness relations follow easily

$$\sum_{r=+,-} u^{(r)}(p) \bar{u}^{(r)}(p) = \frac{(\not{p} + m)}{2m}, \quad \sum_{r=+,-} v^{(r)}(p) \bar{v}^{(r)}(p) = \frac{(\not{p} - m)}{2m} \quad (4.12)$$

where  $\not{p} = \gamma^\mu p_\mu$ . We also have the useful relations:  $m \bar{u}^{(r)}(p) \gamma^\mu u^{(s)}(p) = p^\mu \bar{u}^{(r)}(p) u^{(s)}(p)$  and  $m \bar{v}^{(r)}(p) \gamma^\mu v^{(s)}(p) = -p^\mu \bar{v}^{(r)}(p) v^{(s)}(p)$ .

### 4.3 Fermion propagator

The Fourier transform expansion of  $\psi(x)$  over the complete set of linearly independent plane wave solutions constructed above may be written as

$$\psi(x) = \frac{1}{\sqrt{(2\pi)^3}} \sum_{r=\pm} \int d^2 p^\perp dp^+ \theta(p^+) \sqrt{\frac{m}{p^+}} \left[ b^{(r)}(p) u^{(r)}(p) e^{-ip \cdot x} + d^{\dagger(r)}(p) v^{(r)}(p) e^{ip \cdot x} \right] \quad (4.13)$$

where the  $\theta(p^+)$  is necessarily present. For the dynamical component  $\psi_+ \equiv \Lambda^+ \psi$ , it follows that

$$\psi_+(x) = \frac{\sqrt{\sqrt{2}}}{\sqrt{(2\pi)^3}} \sum_{r=\pm} \int d^2 p^\perp dp^+ \theta(p^+) \left[ b^{(r)}(p) \bar{u}_+^{(r)} e^{-ip \cdot x} + d^{\dagger(r)}(p) \bar{v}_+^{(r)} e^{ip \cdot x} \right]. \quad (4.14)$$

It is straightforward to verify that the anticommutation relations (4.5) for the independent field operator  $\psi_+$  are in fact satisfied if we assume the standard canonical anticommutators, with the nonvanishing ones given by:  $\{b^{(r)}(p), b^{\dagger(s)}(p')\} = \delta_{rs} \delta^2(p^\perp - p'^\perp) \delta(p^+ - p'^+)$  and  $\{d^{(r)}(p), d^{\dagger(s)}(p')\} = \delta_{rs} \delta(p^+ - p'^+) \delta^2(p^\perp - p'^\perp)$ .

The  $\Lambda^+$  projections of our LF spinors are by construction very simple,  $u^{(r)}_+(p) = (\sqrt{2}p^+/m)^{1/2} \Lambda^+ \bar{u}^{(r)}$ ; they are eigenstates of  $\Sigma_3$  as well. This is very convenient since on the LF  $\psi_+$  component is the independent dynamical degrees of freedom while  $\psi_-$  may be eliminated, even in the interacting theory, making use of the constraint equation. The simplified structure of  $\psi_+$  gives rise to appreciable simplifications in the context of LF perturbation theory, compensating to some extent for the nonlinearity of the interaction found along the longitudinal direction  $x^-$ . We have better control [17], say, over recovering the manifest rotational and even Lorentz covariance in the perturbation theory calculations if we use the LF four-spinor introduced above. The propagator for the spinor field  $\psi_+$  also takes a very simple causal form on the LF, resembling the one of the scalar field.

<sup>8</sup>For spin-1/2 case:  $J_j = \Sigma_j/2$ ,  $K_j = i\gamma^0 \gamma^j/2$  where  $j = 1, 2, 3$ .



The free propagator for the independent component  $\psi_+$  in momentum space is easily derived using the above Fourier transform

$$\begin{aligned} \langle 0 | T(\psi_{+A}(x)\psi_{+B}^\dagger(0)) | 0 \rangle &= \\ \langle 0 | \left[ \theta(\tau)\psi_{+A}(x)\psi_{+B}^\dagger(0) - \theta(-\tau)\psi_{+B}^\dagger(0)\psi_{+A}(x) \right] | 0 \rangle &= \\ = \frac{1}{\sqrt{2}} \frac{\Lambda_{AB}^+}{(2\pi)^3} \int d^2 q^\perp dq^+ \theta(q^+) [\theta(\tau)e^{-iqx} - \theta(-\tau)e^{iqx}] & \end{aligned} \quad (4.15)$$

where  $A, B = 1, 2, 3, 4$  label the spinor components. The only relevant differences, compared with the case of the scalar field, are, apart from the appearance of the projection operator, the absence of the factor  $(1/2q^+)$  in the integrand, and the negative sign of the second term in the fermionic case. They, however, compensate and the standard manipulations to factor out the exponential give rise to the factor  $[\theta(q^+) + \theta(-q^+)]$  which may be interpreted as unity in the distribution theory sense, parallel to what we find in the derivation of the scalar field propagator on the LF. Hence

$$\langle 0 | T(\psi_+(x)\psi_+^\dagger(0)) | 0 \rangle = \frac{i}{(2\pi)^4} \int d^4 q \frac{\sqrt{2}q^+ \Lambda^+}{(q^2 - m^2 + i\epsilon)} e^{-iq \cdot x}. \quad (4.16)$$

It may also be derived by functional integral method; we do have to take care of the second class constraint in the measure. The fermionic propagator here contains no instantaneous term usually encountered when doing the old fashioned perturbation theory and the integrand factor may also be expressed as  $\approx [\Lambda^+(\not{q} + m)\Lambda^-/(q^2 - m^2 + i\epsilon)] \gamma^0$ . We verify that the propagator satisfies the equation for the Green's function corresponding to the equation of motion of  $\psi_+$ .

The momentum space representations of the currents and the components of the energy-momentum tensor are derived straightforwardly and they support the usual interpretation of  $b^{\dagger(r)}(p)b^{(r)}(p)$  and  $d^{\dagger(r)}(p)d^{(r)}(p)$  as the number operators. For example, for the canonical Hamiltonian we find

$$\begin{aligned} H_c^{lf} &= \frac{1}{\sqrt{2}} \int d^2 x^\perp dx^- : \psi_+^\dagger (m^2 - \partial_\perp \partial_\perp) \frac{1}{i\partial_-} \psi_+ : \\ &= \sum_{r,s} \int d^3 p d^3 k \theta(p^+) \theta(k^+) : \left[ b^{\dagger(r)}(p) b^{(s)}(k) \tilde{u}_+^{\dagger(r)} \tilde{u}_+^{(s)} \right. \\ &\quad \left. - d^{(r)}(p) d^{\dagger(s)}(k) \tilde{v}_+^{\dagger(r)} \tilde{v}_+^{(s)} \right] : \frac{(m^2 + p^\perp p^\perp)}{2p^+} \delta^3(p - k) \\ &= \sum_r \int d^3 p \theta(p^+) \left[ b^{\dagger(r)}(p) b^{(r)}(p) + d^{\dagger(r)}(p) d^{(r)}(p) \right] \frac{(m^2 + p^\perp p^\perp)}{2p^+} \end{aligned} \quad (4.17)$$

where we use  $\tilde{u}_+^{\dagger(r)} \tilde{u}_+^{(s)} = \tilde{v}_+^{\dagger(r)} \tilde{v}_+^{(s)} = \delta_{rs}/2$ ,  $d^3 p \equiv d^2 p^\perp dp^+$ , and  $: :$  indicates the normal ordering.

#### 4.4 $\Gamma_5$ Symmetry. Chirality transformation on the LF

The  $\gamma_5$  transformation [52],  $\psi \rightarrow \gamma_5 \psi$  on the spinor field is associated with the mass reversal in the Dirac equation. It leaves the Dirac equation form invariant only when the mass is vanishing. On the LF we can construct a generalized  $\Gamma_5$  transformation which restores the form invariance even for the massive field.

Consider the covariant vector and axial current densities defined by  $j^\mu = \bar{\psi} \gamma^\mu \psi$  and  $j_5^\mu = \bar{\psi} \gamma^\mu \gamma_5 \psi$  respectively. The corresponding charge densities are defined on the LF by the  $+$  components of the currents

$$\begin{aligned} j^+ &= \bar{\psi} \gamma^+ \psi = \sqrt{2} \psi_+^\dagger \psi_+ \\ j_5^+ &= \bar{\psi} \gamma^+ \gamma_5 \psi = \sqrt{2} \psi_+^\dagger \Sigma_3 \psi_+. \end{aligned} \quad (4.18)$$

The momentum space representations of the charges are easily derived

$$\begin{aligned} Q &= \int d^2 x^\perp dx^- : j^+ := \sum_r \int d^3 p \theta(p^+) \left[ b^{\dagger(r)}(p) b^{(r)}(p) - d^{\dagger(r)}(p) d^{(r)}(p) \right] \\ Q_5 &= \int d^2 x^\perp dx^- : j_5^+ := \sum_r \int d^3 p \theta(p^+) (r) \left[ b^{\dagger(r)}(p) b^{(r)}(p) + d^{\dagger(r)}(p) d^{(r)}(p) \right] \end{aligned} \quad (4.19)$$

The charges  $Q$  and  $Q_5$  commute with the LF Hamiltonian and are thus constants of motion. The former counts the fermionic number while the latter the twice the projection along the 3-axis of the LF spin operator  $\mathcal{J}_3(p)$  discussed above.

From the commutation relations of the field  $\psi_+$  we derive [9]

$$\begin{aligned} \{\psi_+, Q\} &= \psi_+, \\ \{\psi_-, Q\} &= \psi_-, \\ \{\psi_+, Q_5\} &= \gamma_5 \psi_+ = \Lambda^+ \gamma_5 \psi_+, \\ \{\psi_-, Q_5\} &= \Lambda^- \frac{1}{2i\partial_-} (i\gamma^\perp \partial_\perp + m) \gamma^+ (\gamma_5 \psi_+) \neq \gamma_5 \psi_-. \end{aligned} \quad (4.20)$$

The action of the infinitesimal generators on  $\psi$  is

$$\begin{aligned} \delta_Q \psi = \{\psi, i\epsilon Q\} &= i\epsilon \psi, \\ \delta_{Q_5} \psi = \{\psi, i\epsilon Q_5\} &= i\epsilon \gamma_5 \left[ I - \frac{m}{i\partial_-} \gamma^+ \right] \psi, \end{aligned} \quad (4.21)$$

where we use (4.3) and (4.4). It is well known that the infinitesimal transformation with respect to  $Q$  is associated with the *form invariance* of the Dirac equation  $(i\gamma^\mu \partial_\mu - m)\psi = 0$  and its conjugate under the global phase transformations. This symmetry gives rise to the on shell conserved Noether vector current  $j^\mu$ .

The Dirac equation is form invariant under the  $\gamma_5$  (or chiral transformations) only for the massless theory, when the axial current is also conserved at the classical level. Our discussion on the LF in the Hamiltonian formulation indicates that the Dirac equation is also form invariant under the following nonlocal  $\Gamma_5$  transformation, defined by  $\Gamma_5$

$$\begin{aligned} \psi &\rightarrow \Gamma_5 \psi, \\ \Gamma_5 &= \gamma_5 \left[ I - \frac{m}{i\partial_-} \gamma^+ \right]. \end{aligned} \quad (4.22)$$

This can be demonstrated, say, if we use of the (on shell) identity

$$(i\gamma^\mu \partial_\mu - m) \gamma_5 \left[ I - \frac{m}{i\partial_-} \gamma^+ \right] = -\gamma_5 \left[ I + \frac{m}{i\partial_-} \gamma^+ \right] (i\gamma^\mu \partial_\mu - m). \quad (4.23)$$

The on shell conserved current associated with the  $\Gamma_5$  symmetry, which holds for both the massive and massless fermions, is hence given by

$$\begin{aligned} J_5^\mu &= \bar{\psi} \gamma^\mu \Gamma_5 \psi = j_5^\mu - m \bar{\psi} \gamma^\mu \gamma_5 \gamma^+ \frac{1}{i\partial_-} \psi, \\ \partial_\mu J^\mu &\stackrel{o}{=} 0, \\ J_5^+ &= j_5^+. \end{aligned} \quad (4.24)$$

The chiral charge associated with the  $\Gamma_5$  symmetry coincides with  $Q_5$  and the generalized chiral transformation is  $\psi \rightarrow e^{i\alpha \Gamma_5} \psi$ .

#### 4.5 Helicity Operator, LF Majorana and Weyl fermions

The Fourier transform of the self-charge conjugate *Majorana spinor field* satisfying,  $\psi_M(x) = \psi_{M^c}(x)$ , follows easily from (4.13)

$$\begin{aligned}\psi_M(x) &= \frac{1}{\sqrt{2}}(\psi(x) + \psi_c) \\ &= \frac{1}{\sqrt{(2\pi)^3}} \sum_{r=\pm} \int d^2p^\perp dp^+ \theta(p^+) \sqrt{\frac{m}{p^+}} \left[ b_M^{(r)}(p) u^{(r)}(p) e^{-ip \cdot x} \right. \\ &\quad \left. + b_M^{\dagger(r)}(p) v^{(r)}(p) e^{ip \cdot x} \right]\end{aligned}\quad (4.25)$$

where  $b_M^{(r)}(p) = (b^{(r)}(p) + d^{(r)}(p))/\sqrt{2}$  and the nonvanishing anti-commutator is given by  $\{b_M^{(r)}(p), b_M^{\dagger(s)}(k)\} = \delta^{rs} \delta^3(p - k)$ .

The *chiral* or  $\gamma_5$ -projections of the LF spinor are shown to satisfy the following properties ( $r \gamma_5 u^{(r)}(p; m) = u^{(r)}(p; -m)$  and  $-r \gamma_5 v^{(r)}(p; m) = v^{(r)}(p; -m)$ )

$$\begin{aligned}\frac{(I + r\gamma_5)}{2} u^{(r)}(p) &= N(p) \left[ \sqrt{2} p^+ \Lambda^+ + (\gamma^\perp p_\perp) \Lambda^- \right] \tilde{u}^{(r)} \\ \frac{(I - r\gamma_5)}{2} u^{(r)}(p) &= N(p) m \Lambda^- \tilde{u}^{(r)} \rightarrow 0 \quad \text{for } m \rightarrow 0 \\ \frac{(I - r\gamma_5)}{2} v^{(r)}(p) &= N(p) \left[ \sqrt{2} p^+ \Lambda^+ - (\gamma^\perp p_\perp) \Lambda^- \right] \tilde{v}^{(r)} \\ \frac{(I + r\gamma_5)}{2} v^{(r)}(p) &= N(p) m \Lambda^- \tilde{v}^{(r)} \rightarrow 0 \quad \text{for } m \rightarrow 0\end{aligned}\quad (4.26)$$

along with

$$\begin{aligned}\gamma^\mu p_\mu \left[ \frac{(I + r\gamma_5)}{2} u^{(r)}(p) \right] &= N(p) m^2 \Lambda^- \tilde{u}^{(r)} \rightarrow 0 \quad \text{for } m \rightarrow 0 \\ \gamma_5 \left[ \frac{(I + r\gamma_5)}{2} u^{(r)}(p) \right] &= r \left[ \frac{(I + r\gamma_5)}{2} u^{(r)}(p) \right]\end{aligned}\quad (4.27)$$

etc., and we note that  $[\mathcal{J}_3(p), \gamma_5] = 0$ . In the *massless limit*,  $m \rightarrow 0$ , the projections  $(I \mp \gamma_5) u^{(\pm)}(p)$ ,  $(I \pm \gamma_5) v^{(\pm)}(p)$  vanish. Also, for example, the nonvanishing one

$$\frac{(I + \gamma_5)}{2} u^{(+)}(p) \quad (4.28)$$

is an eigenstate of  $\gamma_5$  and  $\mathcal{J}_3(p)$  with the eigenvalues 1 and 1/2 respectively, while the other one

$$\frac{(I - \gamma_5)}{2} u^{(-)}(p) \quad (4.29)$$

has the corresponding eigenvalues given by -1 and -1/2. The explicit discussion here shows that on the LF the definition of the spin operator for the massive and massless cases gets unified.

The *Helicity operator*  $\hat{h}$  is defined by

$$\hat{h} = \frac{1}{2} \frac{\vec{\Sigma} \cdot \hat{\vec{P}}}{|\hat{\vec{P}}|} = \frac{[\Sigma_3 \hat{P}^3 + \gamma^\perp \hat{P}_\perp \gamma^0 \gamma_5]}{2|\hat{\vec{P}}|} \quad (4.30)$$

which is *not* the same as the LF spin operator.

For *massless fermions* it is easily shown that

$$\begin{aligned}\hat{h}(p) u^{(r)}(p) &= \left(\frac{r}{2}\right) u^{(r)}(p) \\ \hat{h}(p) v^{(r)}(p) &= -\left(\frac{r}{2}\right) v^{(r)}(p)\end{aligned}\quad (4.31)$$

Experimental observations show that only the negative chirality,  $(I - \gamma_5)u^{(-)}(p)/2$ , neutrinos exist. Neutrinos have helicity  $-1/2$ , antineutrinos helicity  $1/2$ . There is no charge conjugation invariance if neutrinos have a definite chirality. The CP transformations of these spinors can be discussed as usual (Appendix B). The normalization factor in the massless case has to be redefined. The massive particle does not have Lorentz invariant helicity; in the rest frame of the particle there is no preferred direction in what to measure spin.

## 4.6 Bilocal operators

From the anticommutators in Sec. 4.1 we may derive the (free theory) equal- $\tau$  current commutation relations, for example,  $[j^+(x), j^+(y)]_\tau = 0$ . The commutators among the other components are derived straightforwardly. They involve bilocal operators [53] of the form  $\bar{\psi}(x)\Gamma\psi(y)$ , with the nonlocality *only* along the longitudinal direction. In the context of the deep inelastic scattering limit they are found relevant in the hadron tensor  $W^{\mu\nu}$  and the explanation of the Bjorken scaling and the introduction of the parton model of Feynman. Similar bilocal operators appear also in bosonic theories, for example, in the LF quantization [54] of Chern-Simons systems. We recall (Sec. 1) that on the LF nonlocality in the  $x^-$  direction does not conflict with the microcausality principle. The bilocals have also been shown useful recently, for example, in the context [55] of the dynamics of hadrons in two dimensions and in revealing the string like structure in  $QCD_2$ .

## 5 LF quantization of Gauge theory

In perturbative QCD we employ, in the interaction representation, the free abelian gauge theory propagator. It is customary on the LF to adopt the light-cone gauge<sup>9</sup>  $A_- = 0$  which results in a simplified interaction Hamiltonian. The noncovariant gauge, however, introduces in the theory undesirable features. The rotational invariance becomes very difficult to track down making the comparison with the conventional theory results sometimes extremely difficult. In the frequently employed old fashioned perturbation theory computations it is sometimes not easy to see if the conventional and the *front form* theories are really in agreement [56]. The LF quantized QCD was recently studied [17] in covariant gauges in the context of the Dyson-Wick perturbation theory expansion based on the LF-time ordered Wick products. Here all the relevant propagators become causal and the rotational invariance is easily recovered, when the LF spinor (4.7) introduced in the Sec. 4 is employed. The loop integrals can also be converted [18] to the Euclidean space integrals and the dimensional regularization may be used.

The Lagrangian density for the Abelian gauge theory written in LF coordinates is

$$\frac{1}{2} [(F_{+-})^2 - (F_{12})^2 + 2F_{+\perp}F_{-\perp}] + B(\partial_+A_- + \partial_-A_+ + \partial_\perp A^\perp) + \frac{\xi}{2}B^2, \quad (5.1)$$

where  $F_{\mu\nu} \equiv (\partial_\mu A_\nu - \partial_\nu A_\mu)$ . The covariant gauge-fixing is introduced by adding to the Lagrangian the linear gauge-fixing term  $B\partial_\mu A^\mu + (\xi/2)B^2$  where  $B$  is the Nakanishi-Lautrup auxiliary field and  $\xi$  is a parameter. The canonical momenta are  $\pi^+ = 0$ ,  $\pi_B = 0$ ,  $\pi^\perp = F_{-\perp}$ ,  $\pi^- = F_{+-} + B$  and the canonical Hamiltonian density is found to be

$$\mathcal{H}_c = \frac{1}{2}(\pi^-)^2 + \frac{1}{2}(F_{12})^2 - A_+(\partial_- \pi^- + \partial_\perp \pi^\perp - 2\partial_- B) - B(\pi^- + \partial_\perp A^\perp) + \frac{1}{2}(1 - \xi)B^2 \quad (5.2)$$

Following the Dirac procedure, the primary constraints are  $\pi^+ \approx 0$ ,  $\pi_B \approx 0$  and  $\eta \equiv \pi^\perp - \partial_- A_\perp + \partial_\perp A_- \approx 0$ , where  $\perp = 1, 2$  and  $\approx$  stands for *weak equality* relation. We now require the persistency in  $\tau$  of these constraints employing the preliminary Hamiltonian, which is obtained by adding to the canonical Hamiltonian the primary constraints multiplied by the Lagrange multiplier fields. We

<sup>9</sup>See the discussion below on the LF quantized two dimensional SM where this gauge is not convenient to employ if we are seeking for nonperturbative effects in the theory.

assume the standard Poisson brackets for the dynamical variables in the computation for obtaining the Hamilton's equations of motion. We are led to the following two secondary constraints

$$\begin{aligned}\Phi &\equiv \partial_- \pi^- + \partial_\perp \pi^\perp - 2\partial_- B \approx 0, \\ \Psi &\equiv \pi^- + 2\partial_- A_+ + \partial_\perp A^\perp - (1 - \xi)B \approx 0.\end{aligned}\quad (5.3)$$

The Hamiltonian is next enlarged by including these additional constraints as well. The procedure is repeated. No more constraints are seen to arise. we now go over from the standard Poisson brackets to the Dirac brackets, such that inside them we are able to substitute the above constraints as *strong* equality. The equal- $\tau$  Dirac bracket  $\{f(x), g(y)\}_D$  which carries this property is constructed straightforwardly. Hamilton's equations now employ the Dirac brackets and the phase space constraints  $\pi^+ = 0$ ,  $\pi_B = 0$ ,  $\eta = 0$ ,  $\Phi = 0$ , and  $\Psi = 0$  then effectively reduce the (extended) Hamiltonian. In the covariant *Feynman gauge* with  $\xi = 1$  the free Hamiltonian takes the simple form

$$H_0^{LF} = -\frac{1}{2} \int d^2 x^\perp dx^- g^{\mu\nu} A_\mu \partial^\perp \partial_\perp A_\nu. \quad (5.4)$$

The theory is canonically quantized through the correspondence  $i\{f(x), g(y)\}_D \rightarrow [f(x), g(y)]$ , the commutator among the corresponding operators.

The equal- $\tau$  commutators of the gauge field are found to be

$$[A_\mu(x), A_\nu(y)]_{x^+=y^+=\tau} = -ig_{\mu\nu} K(x, y) \quad (5.5)$$

where  $K(x, y) = -(1/4)\epsilon(x^- - y^-)\delta^2(x^\perp - y^\perp)$  is nonlocal in the longitudinal coordinate. The transverse components of the gauge field have the physical LF commutators  $[A_\perp(x), A_\perp(y)]_\tau = i\delta_{\perp,\perp'} K(x, y)$ , while for the  $\pm$  components we have only the mixed commutator nonvanishing  $[A_+(x), A_-(y)]_\tau = -iK(x, y)$ , it has a negative sign which indicates the presence of unphysical degrees of freedom in Feynman gauge. For  $\xi \neq 1$  the commutator, for example, of  $A_\pm$  with  $A_\perp$  is found to be nonvanishing. We note that the dimension of the gauge field is  $[A_\mu] = 1/L_\perp$ .

From the discussion analogous to that given in Sec. (2.6) for the scalar field it is clear, from the primary constraints, e.g.,  $\chi^\perp \approx 0$ , in the discussion here, that there are also first class constraints present in the gauge theory. They may be taken care of like in the case of the scalar theory. In the context of perturbation theory we may possibly ignore the zero-longitudinal-mode of the components of the gauge field. However, when dealing with nonperturbative effects they may not be ignored. For example, in the discussion of the (nonperturbative) vacuum structure of the completely soluble  $QED_2$  (SM) theory the zero-momentum mode of  $A_-$  plays a crucial role together with the bosonic condensate variable (Sec. 6).

The Heisenberg equations of motion lead to  $\partial^2 A_\mu = 0$  for all the components, and consequently the Fourier transform of the free gauge field over the complete set of plane wave solutions takes the following form on the LF

$$A^\mu(x) = \frac{1}{\sqrt{(2\pi)^3}} \int d^2 k^\perp dk^+ \frac{\theta(k^+)}{\sqrt{2k^+}} e^{\mu(\lambda)}(k) \left[ a_{(\lambda)}(k^+, k^\perp) e^{-ik \cdot x} + a_{(\lambda)}^\dagger(k^+, k^\perp) e^{ik \cdot x} \right] \quad (5.6)$$

where  $e^{\mu(\lambda)}(k)$ ,  $\lambda = -, +, 1, 2$  label the set of four linearly independent polarization four-vectors.

In the *front form* theory the two transverse (physical) polarization vector are space-like as usual while<sup>10</sup> the other two are null four-vectors. For a fixed  $k^\mu = (k^0, \vec{k})$ , where  $k^0 = |\vec{k}|$ , we may construct them as follows

$$e^{(+)} = (1, \vec{k}/k^0)/\sqrt{2}, \quad e^{(-)} = (1, -\vec{k}/k^0)/\sqrt{2}, \quad e^{(1)} = (0, \vec{\epsilon}(k; 1)), \quad e^{(2)} = (0, \vec{\epsilon}(k; 2)). \quad (5.7)$$

Here (0, 1, 2, 3) components are specified for convenience while  $\vec{\epsilon}(k; 1)$ ,  $\vec{\epsilon}(k; 2)$  and  $\vec{k}/|\vec{k}|$  constitute the usual orthonormal set of 3-vectors with the associated completeness relation. The polarization vectors are orthonormal:  $g_{\mu\nu} e^{(\lambda)\mu}(k) e^{(\sigma)\nu}(k) = g^{\lambda\sigma}$  and satisfy the completeness relation:  $g_{\lambda\sigma} e^{(\lambda)}_\mu(k) e^{(\sigma)}_\nu(k) = g_{\mu\nu}$ .

<sup>10</sup> $e^{(-)}(k)$  is called the dual of  $e^{(+)}(k)$ . Such a pair of null vectors is employed also in the well known ML prescription in the light-cone gauge and in the context of CNPA [20, 21].

The field commutation relations for the gauge field found above are shown to be satisfied if we assume, parallel to the discussion in the fermionic case, the canonical commutation relations:  $[a_{(\lambda)}(k^+, k^\perp), a_{(\sigma)}^\dagger(k'^+, k'^\perp)] = -g_{\lambda\sigma} \delta(k^+ - k'^+) \delta^2(k^\perp - k'^\perp)$ . We note that the operators  $a_{(0)} = (a_{(+)} + a_{(-)})/\sqrt{2}$  and  $a_{(3)} = (a_{(+)} - a_{(-)})/\sqrt{2}$  obey the usual canonical commutation relations except that in the case of  $a_{(0)}$  a negative sign is obtained. The discussion of the Gupta-Bleuler consistency condition then becomes parallel to that in the conventional equal-time treatment of the theory.

The Feynman gauge free gauge field propagator on the LF can be derived straightforwardly

$$\langle 0|T(A_\mu(x)A_\nu(0))|0\rangle = \frac{i}{(2\pi)^4} \int d^4k e^{-ik \cdot x} \frac{-g_{\mu\nu}}{k^2 + i\epsilon} \quad (5.8)$$

The momentum space representations of the components of the energy-momentum tensor are straightforward to derive as in the fermionic case. The canonical Hamiltonian, for example, gets contributions from the physical transversely polarized photons as well as from the longitudinally polarized ones. The Gupta-Bleuler consistency condition is required [17] to be imposed in order to define the physical Hilbert space.

The computations done [17], employing the covariant gauge on the LF, for the electron self-energy, electron-muon scattering, and the Compton scattering demonstrate complete agreement with the results known in the conventional equal-time theory. We find that on the LF the tree level *seagull* term dominates the (classical) Thomas formula for the scattering at vanishingly small photon energies. It is suggestive that on the LF the (conventional theory) semi-classical approximation may reveal itself already at the tree level (after having removed the constraints). We will consider the LF quantized QCD after the study in the *front form* theory of the nonperturbative vacuum structures in some two dimensional completely solvable gauge theories.

## 6 Vacuum Structures in Schwinger and Chiral Schwinger Models

It is pertinent to study two dimensional gauge theories on the LF. The models like SM and CSM can be solved completely. They may give clues, for example, on the accessibility or not, in the fully interacting theory, of certain gauge-fixing condition, found practical in the context of perturbation theory. The study [15] of the SM, for example, shows that the light-cone gauge,  $A_- = 0$ , is *not* convenient on the LF; it would subtract out the gauge invariant information from the theory itself, which is needed for describing the nonperturbative vacuum structure in the theory.

The models mentioned above are known to have non-trivial vacuum structure, a non-perturbative effect, from the studies [57] in the conventional framework. Their study would indicate as to how to look for such and other nonperturbative effects in the LF quantized QCD in 3 + 1 dimensions.

The massless  $QED_2$  or SM is describe by

$$\mathcal{L} = \bar{\psi} i\gamma^\mu \partial_\mu \psi - \frac{1}{4} F^{\mu\nu} F_{\mu\nu} - e \bar{\psi} \gamma^\mu \psi A_\mu. \quad (6.1)$$

Its exact solvability [38] derives from the remarkable property of one-dimensional fermion systems, viz, that they can equivalently be described in terms of canonical one-dimensional boson fields. Some of the correspondences in the abelian bosonization are  $\bar{\psi}\psi = K : \cos 2\sqrt{\pi}\phi :$ ,  $\bar{\psi}\gamma_5\psi = K : \sin 2\sqrt{\pi}\phi :$ ,  $\bar{\psi}\gamma_5\gamma_\mu\psi = \partial_\mu\phi/\sqrt{\pi}$ ,  $\bar{\psi}\gamma_\mu\psi = \epsilon_{\mu\nu}\partial^\nu\phi/\sqrt{\pi}$ ,  $\bar{\psi}i\gamma_\mu\partial^\mu\psi = \frac{1}{2}\partial_\mu\phi\partial^\mu\phi$  where  $\phi$  is a bosonic scalar field and  $K$  is a constant. The fermionic condensate  $\langle \bar{\psi}\psi \rangle_0$ , for example, may then be expressed in terms of the value of the bosonic condensate. The bosonized theory can also be constructed with the use of the functional integral method. The original fermionic and the bosonized theories are *equivalent* in the sense that they have the same current commutation relations and the energy-momentum tensor is the same when expressed in terms of the currents.

For studying nonperturbative vacuum structure the bosonized theory is convenient to use. The *bosonized* version of  $QED_2$  is found to be

$$\mathcal{L} = \frac{1}{2} \partial_\mu \phi \partial^\mu \phi - g A_\mu \epsilon^{\mu\nu} \partial_\nu \phi - \frac{1}{4} F^{\mu\nu} F_{\mu\nu}, \quad (6.2)$$

where  $g = e/\sqrt{\pi}$ . It carries in it all the symmetries of the original fermionic model including the information on the dynamical mass generation [38] for the gauge field. Under the  $U(1)$  gauge field transformation the scalar field is invariant (or neutral) while under the chiral transformations,  $U_5(1)$ , in view of the correspondences above, the field suffers a translation by a constant.

Following the procedure of Secs. 2 and 3 we make the *separation*, of the condensate variable in the scalar field :  $\phi(\tau, x^-) = \omega(\tau) + \varphi(\tau, x^-)$ . The *chiral transformation* is now defined by:  $\omega \rightarrow \omega + \text{const.}$ ,  $\varphi \rightarrow \varphi$ , and  $A_\mu \rightarrow A_\mu$  so that the *boundary conditions* at infinity on the quantum fluctuation field  $\varphi$  are kept unaltered under these transformations and the mathematical framework be considered *well posed*. The bosonized Lagrangian written in the LF coordinates reads as is rewritten as

$$L = \int dx^- \left[ \dot{\varphi} \varphi' + g(A_+ \varphi' - A_- \dot{\varphi}) + \frac{1}{2} (\dot{A}_- - A_+' )^2 \right] - g \dot{\omega} h(\tau), \quad (6.3)$$

where  $h(\tau) = \int dx^- A_-(\tau, x^-)$ , an overdot (a prime) indicates the partial derivative with respect to  $\tau$  ( $x^-$ ). We work in the *continuum* and require (on physical considerations) that the relevant fields satisfy the necessary conditions such that their Fourier transforms with respect to the spatial longitudinal coordinate  $x^-$  exist.

The last term in the Lagrangian density shows that the *light-cone gauge*,  $A_- = 0$ , employed often in perturbation theory computations, *may not be appropriate to use in the fully interacting theory*<sup>11</sup>, if we are seeking to study also the nonperturbative effects in the theory. Also the zero-momentum mode of  $A_-$  is a gauge invariant quantity under the boundary conditions assumed. We may, of course, impose different boundary conditions on the fields or add new ingredients in the theory so as to compensate for the elimination of the physical dynamical variable  $h(\tau)$ . A convenient alternative is the local gauge-fixing condition  $\partial_- A_- = 0$ , which is accessible on the phase space. We remove only the nonzero modes of  $A_-$ .

Following the Dirac method to eliminate the constraints in the *front form* theory only the three linearly independent operators survive: the *condensate*  $\omega$ ,  $h(\tau)$ , the zero-momentum-mode of  $A_-$  and canonically conjugate to  $\omega$  as well, and  $\varphi$  which satisfies the LF commutator while it commutes with the others. The  $H^{lf}$  contains in it only the field  $\varphi$ . The Hilbert space can thus be described in two different fashions. Selecting  $\varphi$  and  $h$  as forming the complete set of mutually commuting operators leads to the *chiral vacua* while selecting  $\varphi$  together with  $\omega$  leads to the description built on the *condensate* or  $\theta$ -vacua. In the  $QED_2$  the  $\omega$  is *not* a background field rather it is shown [15] to be an operator and its eigenvalues, with continuous spectrum, label the *condensate* vacua of the theory. The cluster decomposition property requirement [23] indicates the preference in favor of the *condensate* vacua.

The other related gauge theory model is the *chiral QED<sub>2</sub>* or CSM described by

$$\mathcal{L} = -\frac{1}{4} F^{\mu\nu} F_{\mu\nu} + \bar{\psi}_R i \gamma^\mu \partial_\mu \psi_R + \bar{\psi}_L \gamma^\mu (i \partial_\mu + 2e\sqrt{\pi} A_\mu) \psi_L, \quad (6.4)$$

where<sup>12</sup>  $\psi = \psi_R + \psi_L$  is a two-component spinor field and  $A_\mu$  is the abelian gauge field,  $\gamma_5 \psi_L = -\psi_L$ , and  $\gamma_5 \psi_R = \psi_R$ . The classical Lagrangian is invariant under the local  $U(1)$  gauge transformations  $A_\mu \rightarrow A_\mu + \partial_\mu \alpha / (2\sqrt{\pi}e)$ ,  $\psi \rightarrow [P_R + e^{i\alpha} P_L] \psi$  and under the global  $U(1)_5$  chiral transformations  $\psi \rightarrow \exp(i\gamma_5 \alpha) \psi$ .

<sup>11</sup>Similar considerations are clearly pertinent to 3 + 1 dimensional QCD as well.

<sup>12</sup>In two dimensions the  $\pm$  projections of the spinor coincide with the chiral or  $\gamma_5$  projections. We define  $\gamma^0 = \sigma_1$ ,  $\gamma^1 = i\sigma_2$ ,  $\gamma_5 = \gamma^0 \gamma^1 = -\sigma_3$ ,  $\Lambda^- = \gamma^0 \gamma^- / \sqrt{2} = (1 - \gamma_5)/2 \equiv P_L$ ,  $\Lambda^+ = \gamma^0 \gamma^+ / \sqrt{2} = (1 + \gamma_5)/2 \equiv P_R$ ,  $x^\mu : (x^+ \equiv \tau, x^- \equiv x)$  with  $\sqrt{2}x^\pm = \sqrt{2}x_\mp = (x^0 \pm x^1)$ ,  $A^\pm = A_\mp = (A^0 \pm A^1)/\sqrt{2}$ ,  $\psi_{L,R} = P_{L,R} \psi$ ,  $\bar{\psi} = \psi^\dagger \gamma^0$ .

The bosonized version is convenient to study the vacuum structure; it is shown to be given by

$$S = \int d^2x \left[ -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{1}{2}\partial_\mu\phi\partial^\mu\phi + eA_\nu(\eta^{\mu\nu} - \epsilon^{\mu\nu})\partial_\mu\phi + \frac{1}{2}ae^2A_\mu A^\mu \right] \quad (6.5)$$

Here the explicit mass term for the gauge field parametrized by the constant parameter  $a$  represents a regularization ambiguity [58] and the breakdown of  $U(1)$  gauge symmetry. The model has received much attention since Jackiw and Rajaraman [58] pointed out that, despite the gauge anomaly the theory can be shown to be unitary and consistently quantized. In the LF coordinates it reads as

$$S = \int d^2x \left[ \dot{\varphi}\varphi' + \frac{1}{2}(\dot{A}_- - A'_+)^2 + ae^2[A_+ + \frac{2}{ae}(\dot{\omega} + \dot{\varphi})]A_- \right]. \quad (6.6)$$

We note now that  $A_+$  appears in the action as an *auxiliary* field, without a kinetic term. It is clear that the condensate variable may thus be subtracted out from the theory using the frequently adopted procedure of *field redefinition* [16] on it:  $A_+ \rightarrow A_+ - 2\dot{\omega}/(ae)$ , obtaining thereby

$$\mathcal{L}_{CSM} = \dot{\varphi}\varphi' + \frac{1}{2}(\dot{A}_- - A'_+)^2 + 2e\dot{\varphi}A_- + ae^2A_+A_-, \quad (6.7)$$

which signals the emergence of a *different structure* of the Hilbert space compared to that of the SM.

The Lagrange equations in the CSM follow to be

$$\begin{aligned} \partial_+\partial_-\varphi &= -e\partial_+A_-, \\ \partial_+\partial_+A_- - \partial_+\partial_-A_+ &= ae^2A_+ + 2e\partial_+\varphi, \\ \partial_-\partial_-A_+ - \partial_+\partial_-A_- &= ae^2A_-. \end{aligned} \quad (6.8)$$

and for  $a \neq 1$  they lead to:

$$\begin{aligned} \partial^2 G(\tau, x) &= 0 \\ \left[ \partial^2 + \frac{e^2 a^2}{(a-1)} \right] E(\tau, x) &= 0, \end{aligned} \quad (6.9)$$

where  $E = (\partial_+A_- - \partial_-A_+)$  and  $G = (E - ae\varphi)$ . Both the massive and massless scalar excitations are present in the theory and the tachyons would be absent in the spectrum if  $a > 1$ ; the case considered in this paper. We will confirm in the Hamiltonian framework below that the  $E$  and  $G$  represent, in fact, the two independent field operators on the LF phase space.

The Dirac procedure as applied to the very simple action (6.7) of the CSM is straightforward. The canonical momenta are  $\pi^+ \approx 0, \pi^- \equiv E = \dot{A}_- - A'_+, \pi_\varphi = \varphi' + 2eA_-$  which result in two primary weak constraints  $\pi^+ \approx 0$  and  $\Omega_1 \equiv (\pi_\varphi - \varphi' - 2eA_-) \approx 0$ . A secondary constraint  $\Omega_2 \equiv \partial_-E + ae^2A_- \approx 0$  is shown to emerge when we require the  $\tau$  independence (persistency) of  $\pi^+ \approx 0$  employing the preliminary Hamiltonian

$$H' = H_c^{lf} + \int dx u_+ \pi^+ + \int dx u_1 \Omega_1, \quad (6.10)$$

where  $u_+$  and  $u_1$  are the Lagrange multiplier fields and  $H_c^{lf}$  is the canonical Hamiltonian

$$H_c^{lf} = \int dx \left[ \frac{1}{2}E^2 + EA'_+ - ae^2A_+A_- \right]. \quad (6.11)$$

and we assume initially the standard equal- $\tau$  Poisson brackets:  $\{E^\mu(\tau, x^-), A_\nu(\tau, y^-)\} = -\delta_\nu^\mu \delta(x^- - y^-)$ ,  $\{\pi_\varphi(\tau, x^-), \varphi(\tau, y^-)\} = -\delta(x^- - y^-)$  etc.. The persistency requirement for  $\Omega_1$  results in an equation for determining  $u_1$ . The procedure is repeated with the following extended Hamiltonian which includes in it also the secondary constraint

$$H_e^{lf} = H_c^{lf} + \int dx u_+ \pi^+ + \int dx u_1 \Omega_1 + \int dx u_2 \Omega_2. \quad (6.12)$$



No more secondary constraints are seen to arise; we are left with the persistency conditions which determine the multiplier fields  $u_1$  and  $u_2$  while  $u_+$  remains undetermined. We also find<sup>13</sup>  $(C)_{ij} = \{\Omega_i, \Omega_j\} = D_{ij} (-2\partial_x \delta(x-y))$  where  $i, j = 1, 2$  and  $D_{11} = 1$ ,  $D_{22} = ae^2$ ,  $D_{12} = D_{21} = -e$  and that  $\pi^+$  has vanishing brackets with  $\Omega_{1,2}$ . The  $\pi^+ \approx 0$  is first class weak constraint while  $\Omega_1$  and  $\Omega_2$ , which does not depend on  $A_+$  or  $\pi^+$ , are second class ones.

We go over from the Poisson bracket to the Dirac bracket  $\{, \}_D$  constructed in relation to the pair,  $\Omega_1 \approx 0$  and  $\Omega_2 \approx 0$

$$\{f(x), g(y)\}_D = \{f(x), g(y)\} - \int \int dudv \{f(x), \Omega_i(u)\} (C^{-1}(u, v))_{ij} \{\Omega_j(v), g(y)\}. \quad (6.13)$$

Here  $C^{-1}$  is the inverse of  $C$  and we find  $(C^{-1}(x, y))_{ij} = B_{ij} K(x, y)$  with  $B_{11} = a/(a-1)$ ,  $B_{22} = 1/[(a-1)e^2]$ ,  $B_{12} = B_{21} = 1/[(a-1)e]$ , and  $K(x, y) = -\epsilon(x-y)/4$ . Some of the Dirac brackets are  $\{\varphi, \varphi\}_D = B_{11} K(x, y)$ ;  $\{\varphi, E\}_D = eB_{11} K(x, y)$ ;  $\{E, E\}_D = ae^2 B_{11} K(x, y)$ ;  $\{\varphi, A_-\}_D = -B_{12} \delta(x-y)/2$ ;  $\{A_-, E\}_D = B_{11} \delta(x-y)/2$ ;  $\{A_-, A_-\}_D = B_{12} \partial_x \delta(x-y)/2$  and the only nonvanishing one involving  $A_+$  or  $\pi^+$  is  $\{A_+, \pi^+\}_D = \delta(x-y)$ .

The equations of motion employ now the Dirac brackets and inside them, in view of their very construction, we may set  $\Omega_1 = 0$  and  $\Omega_2 = 0$  as strong relations. The Hamiltonian is therefore effectively given by  $H_e$  with the terms involving the multipliers  $u_1$  and  $u_2$  dropped. The multiplier  $u_+$  is not determined since the constraint  $\pi^+ \approx 0$  continues to be first class even when the above Dirac bracket is employed. The variables  $\pi_\varphi$  and  $A_-$  are then removed from the theory leaving behind  $\varphi$ ,  $E$ ,  $A_+$ , and  $\pi^+$  as the remaining independent variables. The canonical Hamiltonian density reduces to  $\mathcal{H}_c^{lf} = E^2/2 + \partial_-(A_+ E)$  while  $\dot{A}_+ = \{A_+, H_e^{lf}\}_D = u_+$ . The surface term in the canonical LF Hamiltonian may be ignored if, say,  $E(= F_{+-})$  vanishes at infinity. The variables  $\pi^+$  and  $A_+$  are then seen to describe a decoupled (from  $\varphi$  and  $E$ ) free theory and we may hence drop these variables as well. The effective LF Hamiltonian thus takes the simple form

$$H_{CSM}^{lf} = \frac{1}{2} \int dx E^2, \quad (6.14)$$

which is to be contrasted with the one found in the conventional treatment [57].  $E$  and  $G$  (or  $E$  and  $\varphi$ ) are now the independent variables on the phase space and the Lagrange equations are verified to be recovered for them, which assures us of the selfconsistency [25]. We stress that in our discussion we do *not* employ any gauge-fixing. The same result for the Hamiltonian could be alternatively obtained<sup>14</sup>, however, if we did introduce the gauge-fixing constraint  $A_+ \approx 0$  and made further modification on  $\{, \}_D$  in order to implement  $A_+ \approx 0, \pi^+ \approx 0$  as well. That it is accessible on the phase space to take care of the remaining first class constraint, but not in the bosonized Lagrangian, follows from the Hamiltons eqns. of motion. We recall [15] that in the SM  $\varphi$ ,  $\omega$ , and  $\pi_\omega = (e/\sqrt{\pi}) \int dx A_-$  were shown to be the independent operators and that the matter field  $\varphi$  appeared instead in the LF Hamiltonian. The *canonical quantization* is performed via the correspondence  $i\{f, g\}_D \rightarrow [f, g]$  and we find the following equal- $\tau$  commutators

$$\begin{aligned} [E(x), E(y)] &= iK(x, y)a^2e^2/(a-1), \\ [G(x), E(y)] &= 0, \\ [G(x), G(y)] &= ia^2e^2K(x, y). \end{aligned} \quad (6.15)$$

For  $a > 1$ , when the tachyons are absent as seen from (6), these commutators are also physical and the independent field operators  $E$  and  $G$  generate the Hilbert space with a tensor product structure of the Fock spaces  $F_E$  and  $F_G$  of these fields with the positive definite metric.

<sup>13</sup>We make the convention that the first variable in an equal- $\tau$  bracket refers to the longitudinal coordinate  $x^- \equiv x$  while the second one to  $y^- \equiv y$  while  $\tau$  is suppressed.

<sup>14</sup>A similar discussion is encountered also in the LF quantized Chern-Simons-Higgs system [54].

The commutators obtained can be realized in the momentum space through the following Fourier transforms

$$\begin{aligned} E(x, \tau) &= \frac{ae}{\sqrt{(a-1)\sqrt{2\pi}}} \int_{-\infty}^{\infty} dk \frac{\theta(k)}{\sqrt{2k}} [d(k, \tau)e^{-ikx} + d^\dagger(k, \tau)e^{ikx}], \\ G(x, \tau) &= \frac{ae}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dk \frac{\theta(k)}{\sqrt{2k}} [g(k, \tau)e^{-ikx} + g^\dagger(k, \tau)e^{ikx}], \end{aligned} \quad (6.16)$$

if the operators  $(d, g, d^\dagger, g^\dagger)$  satisfy the well known canonical commutation relations of two independent harmonic oscillators; the well known set of Schwinger's bosonic oscillators, often employed in the angular momentum theory. The expression for the Hamiltonian becomes

$$H_{CSM}^{lf} = \delta(0) \frac{a^2 e^2}{2(a-1)} \int \frac{dk}{2k} \theta(k) N_d(k, \tau) \quad (6.17)$$

where we have dropped the infinite zero-point energy term and note that [22]  $[d^\dagger(k, \tau), d(l, \tau)] = -\delta(k-l)$ ,  $d^\dagger(k, \tau)d(k, \tau) = \delta(0)N_d(k, \tau)$  etc. with similar expressions for the independent  $g$ -oscillators. We verify that  $[N_d(k, \tau), N_d(l, \tau)] = 0$ ,  $[N_d(k, \tau), N_g(l, \tau)] = 0$ ,  $[N_d(k, \tau), d^\dagger(k, \tau)] = d^\dagger(k, \tau)$  etc..

The Fock space can hence be built on a basis of eigenstates of the hermitian number operators  $N_d$  and  $N_g$ . The ground state of CSM is degenerate and described by  $|0\rangle = |E=0\rangle \otimes |G\rangle$  and it carries vanishing LF energy in agreement with the conventional theory discussion [57]. For a fixed  $k$  these states,  $|E=0\rangle \otimes (g^\dagger(k, \tau)^n / \sqrt{n!})|0\rangle$ , are labelled by the integers  $n = 0, 1, 2, \dots$ . The  $\theta$ -vacua are absent in the CSM. However, we recall [15] that in the SM the degenerate *chiral vacua* are also labelled by such integers. We remark also that on the LF we work in the Minkowski space and that in our discussion we do *not* make use of the Euclidean space theory action, where the (classical) vacuum configurations of the Euclidean theory gauge field, belonging to the distinct topological sectors, are useful, for example, in the functional integral quantization of the gauge theory.

On the LF both the bosonized SM and CSM are described in terms of a minimum number of dynamical variables, which survive after the elimination of the phase space constraints. We recall that the introduction of the *bosonic condensate* variable  $\omega(\tau)$  (or in general  $\omega(\tau, x^\perp)$ ) corresponds to the gauge-fixing required in order to deal with the first class constraint  $\int (\pi - \partial_- \phi) \approx 0$ . On the other hand we have the gauge invariant zero-momentum mode  $h(\tau)$  of the gauge field  $A_-$ , apart from the quantum fluctuation field  $\varphi$ . They are in a sense the minimal set of operators which survive in the *front form* theory. With their help the vacuum structures of both the SM and CSM are described in a very economical and transparent fashion on the LF, which agree with the conventional theory conclusions. In the latter, however, we have to go through quite an elaborate and extensive discussion [57]. Finally, if we did adopt the light-cone gauge we must compensate for the loss of the gauge invariant information by some other ingredient, say, by imposing more complicated boundary conditions on the fields involved or by introducing new fields.

## 7 QCD in Covariant gauges

We describe briefly the recent study [17] done on the *front form* QCD in covariant gauges. The Lagrangian density corresponding to the quantum action of QCD is described in standard notation by

$$\mathcal{L}_{QCD} = -\frac{1}{4} F^{a\mu\nu} F^a_{\mu\nu} + B^a \partial_\mu A^{a\mu} + \frac{\xi}{2} B^a B^a + i \partial^\mu \chi_1^a \mathcal{D}^{ac}_\mu \chi_2^c + \bar{\psi}^i (i \gamma^\mu D^i_\mu - m \delta^{ij}) \psi^j \quad (7.1)$$

Here  $\psi^j$  is the quark field with color index  $j = 1..N_c$  for an  $SU(N_c)$  color group,  $A^a_\mu$  the gluon field,  $F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + g f^{abc} A^b_\mu A^c_\nu$  the field strength,  $\mathcal{D}^{ac}_\mu = (\delta^{ac} \partial_\mu + g f^{abc} A^b_\mu)$ ,  $D^{ij}_\mu \psi^j = (\delta^{ij} \partial_\mu - i g A^a_\mu (\lambda^a/2)^{ij}) \psi^j$ ,  $a = 1..(N_c^2 - 1)$  the gauge group index etc. The covariant gauge-fixing is introduced by adding to the Lagrangian the linear gauge-fixing term  $(B^a \partial_\mu A^{a\mu} + \frac{\xi}{2} B^a B^a)$

where  $B^a$  is the Nakanishi-Lautrup auxiliary field and  $\xi$  is a parameter. The  $\chi_1^a$  and  $\chi_2^a$  are the (hermitian) anticommuting Faddeev-Popov ghost fields, and the action is invariant under the BRS transformation.

The quark field term in LF coordinates reads

$$\begin{aligned} \bar{\psi}^i(i\gamma^\mu D_{\mu}^{ij} - m\delta^{ij})\psi^j &= i\sqrt{2}\bar{\psi}_+^i\gamma^0 D_{+}^{ij}\psi_+^j + \bar{\psi}_+^i(i\gamma^\perp D_{\perp}^{ij} - m\delta^{ij})\psi_-^j \\ &+ \bar{\psi}_-^i[i\sqrt{2}\gamma^0 D_{-}^{ij}\psi_-^j + (i\gamma^\perp D_{\perp}^{ij} - m\delta^{ij})\psi_+^j] \end{aligned} \quad (7.2)$$

where  $D_{\pm}^{ij} = (\delta^{ij}\partial_{\pm} - igA_{\pm}^a(\lambda^a/2)^{ij})$ . It shows that the minus components  $\psi_-^j$  are in fact nondynamical (Lagrange multiplier) fields without kinetic terms. The variation of the action with respect to  $\bar{\psi}_-^j$  and  $\psi_-^j$  leads to the following gauge covariant constraint equation

$$i\sqrt{2}D_{-}^{ij}\psi_-^j = -(i\gamma^0\gamma^\perp D_{\perp}^{ij} - m\gamma^0\delta^{ij})\psi_+^j, \quad (7.3)$$

and its conjugate. The  $\psi_-^j$  components may thus be eliminated in favor of the dynamical components  $\psi_+^j$

$$\psi_-^j(x) = \frac{i}{\sqrt{2}} \left[ U^{-1}(x|A_-) \frac{1}{\partial_-} U(x|A_-) \right]^{jk} (i\gamma^0\gamma^\perp D_{\perp}^{kl} - m\gamma^0\delta^{kl})\psi_+^l(x). \quad (7.4)$$

Here, for a fixed  $\tau$ ,  $U \equiv U(x|A_-)$  is an  $N_c \times N_c$  gauge matrix in the fundamental representation of  $SU(N_c)$  and it satisfies

$$\partial_- U(x|A_-) = -ig U(x|A_-) A_-(x) \quad (7.5)$$

with  $A_- = A_-^a \lambda^a/2$ . It has the formal solution

$$U(x^-, x^\perp|A_-) = U(x_-^0, x^\perp|A_-) \bar{\mathcal{P}} \exp \left\{ -ig \int_{x_-^0}^{x^-} dy^- A_-(y^-, x^\perp) \right\} \quad (7.6)$$

where  $\bar{\mathcal{P}}$  indicates the anti-path-ordering along the longitudinal direction  $x^-$ .  $U$  has a series expansion in the powers of the coupling constant.

The Hamiltonian density in Feynman gauge is

$$\begin{aligned} \mathcal{H}^{LF} &= \mathcal{H}_0 + \mathcal{H}_{int} \\ &= -\frac{1}{2}g^{\mu\nu}A_\mu^a\partial^\perp\partial_\perp A_\nu^a - g\sqrt{2}\bar{\psi}_+^i\gamma^0 A_+^{ij}\psi_+^j \\ &- \bar{\psi}_+^i[\delta^{ij}(i\gamma^\perp\partial_\perp - m) + g\gamma^\perp A_{\perp}^{ij}]\psi_-^j + \frac{g}{2}f^{abc}(\partial_\mu A_\nu^a - \partial_\nu A_\mu^a)A^{b\mu}A^{c\nu} \\ &+ \frac{g^2}{4}f^{abe}f^{cde}A_\mu^a A_\nu^b A^{c\mu}A^{d\nu} + \partial^\mu(\bar{\chi}^a)\partial_\mu\chi^a + gf^{abc}(\partial^\mu\bar{\chi}^a)\chi^b A^c_\mu \end{aligned} \quad (7.7)$$

where  $\psi_-^j$  is given above, we have set  $\sqrt{2}\chi = (\chi_1 + i\chi_2)$ ,  $\sqrt{2}\bar{\chi} = (\chi_1 - i\chi_2)$ , and in  $\mathcal{H}^{LF}$  the cubic and higher order terms belong to  $\mathcal{H}_{int}$  which is also understood to be normal ordered. It is worth remarking that despite the presence of the longitudinal operators  $a_\pm$  and  $a_\pm^\dagger$  in the fields  $A_\mu$ , there are no non-zero matrix elements involving these quanta as external lines in view of the commutation relations of these operators as discussed in the previous section.

The perturbation theory expansion in the interaction representation where we time order with respect to the LF time  $\tau$  is built following the Dyson-Wick procedure. We will illustrate it in our context through some explicit computations, for simplicity, in QED where  $U(x|A_-) = \exp\{-ie \int_{x_-^0}^{x^-} du^- A_-(\tau, u^-, x^\perp)\}$  and  $D_\mu = (\partial_\mu - ieA_\mu)$ . We observe that a *seagull* term of the order  $e^2$  is present in the interaction Hamiltonian at the tree level; like that found also in the scalar field QED.

Towards an illustration consider the computation of *Electron Self-Energy*. The contribution from the longitudinal components arises from

$$e^2 \int d^4 x_1 d^4 x_2 : \psi_+^\dagger(x_2) (m + i \not{\partial}_2^T) \int_{-\infty}^{\infty} \frac{1}{2} dy_2^- \epsilon(x_2^- - y_2^-) \{ \int_{y_2^-}^{x_2^-} du_2^- \dot{A}_-(u_2) \} (m - i \not{\partial}_2^T) \bar{\psi}_+(y_2) \bar{\psi}_+^\dagger(x_1) \psi_+(x_1) \dot{A}_+(x_1) : \quad (7.8)$$

leading to

$$e^2 \int d^4 q \frac{\bar{u}^{(r)}(p) [\gamma^-(m + \not{q}^T) \gamma^+] u^{(s)}(p)}{[(p - q)^2 + i\epsilon](q^2 - m^2 + i\epsilon)} (-g_{+-}) \quad (7.9)$$

The graph with the  $A_+$  and  $A_-$  interchanged gives rise to a similar expression with  $g_{+-} \rightarrow g_{-+}$  while  $\gamma^\pm \rightarrow \gamma^\mp$ . The matrix elements following from the four graphs corresponding to the exchange of the (photon) fields  $A_1$  and  $A_2$  is also written down by simple inspection. As in the earlier case the expressions get simplified in virtue of (10) and acquire the covariant form encountered in the conventional covariant perturbation theory. The complete matrix element is found to be

$$e^2 \int d^4 q \frac{\bar{u}^{(r)}(p) [\gamma^\mu (m + \not{q}) \gamma^\nu] u^{(s)}(p)}{[(p - q)^2 + i\epsilon](q^2 - m^2 + i\epsilon)} (-g_{\mu\nu}) \quad (7.10)$$

where  $\tilde{q}^\mu \equiv ((m^2 + q^\perp q^\perp)/(2q^+), q^+, q^\perp)$  and the integration measure is  $d^4 q = d^2 q^\perp dq^+ dq^-$ . Considering that the integrand has the pole at  $q^2 - m^2 \approx 0$  we may regard the expression obtained [17] on the LF to be effectively identical to the one obtained in the conventional covariant theory framework. The discussion parallel to that given here may be followed also in the context of the light-cone gauge. The latter, however, demands the further introduction [59] of a light-like vector  $n^\mu = (n^0, \vec{n})$  and its dual  $\tilde{n}^\mu = (n^0, -\vec{n})$  in order to evaluate the corresponding Feynman integrals in a consistent manner.

## 8 Conclusions

Collected below are some of the interesting conclusions we seem to reach.

- The LF hyperplane is *equally valid and appropriate* as the conventional equal-time one for the field theory quantization.
- The appearance of the nonlocality along the longitudinal direction in the *front form* quantized theory is not unexpected; it does not conflict with the microcausality (or cluster decomposition) principle.
- The covariant phase space and Fourier expansion considerations based on the description of the relativistic theory using light-cone coordinates lead to the LF commutator for the free scalar field, which is nonlocal in the longitudinal direction.
- The hyperplanes  $x^\pm = 0$  define the characteristic surfaces of a hyperbolic partial differential equation. From the mathematical theory of classical partial differential equations [24] it is known that the Cauchy initial value problem would require us to specify the data on both the hyperplanes. From our studies we conclude [16] that it is sufficient in the *front form* theory to choose one of the two LF hyperplanes for canonically quantizing the theory.
  - In the quantized theory the equal- $\tau$  commutators of the field operators, at a fixed initial LF-time, form now a part of the initial data instead and we deal with operator differential equations.
  - The information on the commutators on the other characteristic hyperplane seems already to be contained [15] in the quantized theory; it may not, in general, be required to specify it separately.

- The constrained phase space dynamics in the LF theory with one more kinematical generator and the inherent symmetry with regard to  $x^\pm$  result in a reduced number of independent field operators. The discussion of the Hilbert space becomes more transparent compared to that in the conventional treatment. The lack of manifest covariance which appears problematic can be handled<sup>15</sup> by employing, for example, the *LF four-spinor* [15] and the Fourier transform of the spinor field as defined [17] in Sec. 4.
- On the LF the  $\gamma_5$  symmetry of free massless Dirac equation can be generalized to a nonlocal (chiral)  $\Gamma_5$  symmetry valid also in the massive case. The Weyl and Majorana spinors and the helicity operator may be defined on the LF in straightforward fashion.
- The zero-longitudinal-momentum modes of the fields are important for describing the non-perturbative effects on the LF. In the scalar and gauge theories they are dynamical variables in the frame work of the *standard* Dirac procedure. The *separation*  $\phi(\tau, x^-, x^\perp) = \omega(\tau, x^\perp) + \varphi(\tau, x^-, x^\perp)$  introduced in Secs. 2.6, 5 correspond to the gauge-fixing conditions [25] required to be introduced in the theory for handling first class constraints.
  - In the case of the scalar theory we obtain constraint equations which enable us to describe SSB and the (tree level) Higgs mechanism. Associated to the local theory in the conventional coordinates we find a nonlocal LF Hamiltonian.
  - The gauge field zero modes play a crucial role in the description of the nonperturbative vacuum structures in the LF quantized SM and CSM. They also indicate that the (popular) light-cone gauge may not be accessible in the *front form* theory if we are concerned with the study of nonperturbative effects.
- The physical content following from the *front form* theory is the same, even though arrived at through different description on the LF, when compared with the one in the *instant form* case.
- Not all the constraints in the LF theory need to be solved first before considering its renormalization; it is sometimes convenient to obtain some of them as renormalized constraint equations [18] instead.
- In the conventional treatment we may be required to introduce external constraints in the quantized theory based on physical considerations, say, while describing the SSB. The analogous relevant constraints in the *front form* theory appear to be already contained in the quantized theory.
- On the LF the quantized theory of *chiral boson* appears straightforward (Sec. 3.4). The field commutator does not conflict with the microcausality principle.
- A theoretical demonstration of the well accepted notion that a classical model field theory must be upgraded first through its quantization before we confront it with the experimental data, seems to emerge.
- The LF quantized QCD employing covariant gauges [17] looks promising. All of the propagators become causal and the covariance of the theory is tractable. The semiclassical theory is found revealed at the tree level. The algebra of bilocals in the LF quantized theory may help reveal the string like structure as seems to be found [55], for example, in  $QCD_2$ .
- The recently proposed BRS-BFT [60] quantization procedure is extended straightforwardly on the LF (Appendix C).
- It is well known that topological considerations are often required in the field theory quantization employing the functional integral method, where the Euclidean theory action is employed. The corresponding ingredients seem to arise in the canonically quantized *front*

---

<sup>15</sup>See also, [54]

form theory as well but with different interpretation. This is suggested, for example, from the studies of the SM, CSM, and the study of the *kink* solutions.

- In connection with the relativistic bound state problem, not touched upon in this article, the LF Tamm-Dancoff method [9, 56] and Bethe-Salpeter dynamics on the covariant null plane [19, 20, 21] seem to be promising alternatives to lattice gauge theory approach.

## Acknowledgements

The author acknowledges with thanks the helpful comments from Stan Brodsky, Richard Blankenbecler and Sidney Drell. The hospitality offered to him at the SLAC and a financial grant of Prociência program of the UERJ, Rio de Janeiro, Brasil, are gratefully acknowledged.

## Appendix A: Poincaré Generators on the LF

The Poincaré generators in coordinate system  $(x^0, x^1, x^2, x^3)$ , satisfy  $[M_{\mu\nu}, P_\sigma] = -i(P_\mu g_{\nu\sigma} - P_\nu g_{\mu\sigma})$  and  $[M_{\mu\nu}, M_{\rho\sigma}] = i(M_{\mu\rho}g_{\nu\sigma} + M_{\nu\sigma}g_{\mu\rho} - M_{\nu\rho}g_{\mu\sigma} - M_{\mu\sigma}g_{\nu\rho})$  where the metric is  $g_{\mu\nu} = \text{diag}(1, -1, -1, -1)$ ,  $\mu = (0, 1, 2, 3)$  and we take  $\epsilon_{0123} = \epsilon_{-+12} = 1$ . If we define  $J_i = -(1/2)\epsilon_{ikl}M^{kl}$  and  $K_i = M_{0i}$ , where  $i, j, k, l = 1, 2, 3$ , we find  $[J_i, F_j] = i\epsilon_{ijk}F_k$  for  $F_l = J_l, P_l$  or  $K_l$  while  $[K_i, K_j] = -i\epsilon_{ijk}J_k$ ,  $[K_i, P_l] = -iP_0g_{il}$ ,  $[K_i, P_0] = iP_i$ , and  $[J_i, P_0] = 0$ .

The LF generators are  $P_+, P_-, P_1, P_2, M_{12} = -J_3, M_{+-} = -K_3, M_{1-} = -(K_1 + J_2)/\sqrt{2} \equiv -B_1, M_{2-} = -(K_2 - J_1)/\sqrt{2} \equiv -B_2, M_{1+} = -(K_1 - J_2)/\sqrt{2} \equiv -S_1$  and  $M_{2+} = -(K_2 + J_1)/\sqrt{2} \equiv -S_2$ . We find  $[B_1, B_2] = 0$ ,  $[B_a, J_3] = -i\epsilon_{ab}B_b$ ,  $[B_a, K_3] = iB_a$ ,  $[J_3, K_3] = 0$ ,  $[S_1, S_2] = 0$ ,  $[S_a, J_3] = -i\epsilon_{ab}S_b$ ,  $[S_a, K_3] = -iS_a$  where  $a, b = 1, 2$  and  $\epsilon_{12} = -\epsilon_{21} = 1$ . Also  $[B_1, P_1] = [B_2, P_2] = iP^+$ ,  $[B_1, P_2] = [B_2, P_1] = 0$ ,  $[B_a, P^-] = iP_a$ ,  $[B_a, P^+] = 0$ ,  $[S_1, P_1] = [S_2, P_2] = iP^-$ ,  $[S_1, P_2] = [S_2, P_1] = 0$ ,  $[S_a, P^+] = iP_a$ ,  $[S_a, P^-] = 0$ ,  $[B_1, S_2] = -[B_2, S_1] = -iJ_3$ ,  $[B_1, S_1] = [B_2, S_2] = -iK_3$ . For  $P_\mu = i\partial_\mu$ , and  $M_{\mu\nu} \rightarrow L_{\mu\nu} = i(x_\mu\partial_\nu - x_\nu\partial_\mu)$  we find  $B_a = (x^+P^a - x^aP^+)$ ,  $S_a = (x^-P^a - x^aP^-)$ ,  $K_3 = (x^-P^+ - x^+P^-)$  and  $J_3 = (x^1P^2 - x^2P^1)$ . Under the conventional *parity* operation  $\mathcal{P}$ : ( $x^\pm \leftrightarrow x^\mp, x^{1,2} \rightarrow -x^{1,2}$ ) and ( $p^\pm \leftrightarrow p^\mp, p^{1,2} \rightarrow -p^{1,2}$ ), we find  $\vec{J} \rightarrow \vec{J}, \vec{K} \rightarrow -\vec{K}, B_a \rightarrow -S_a$  etc.. The six generators  $P_i, M_{kl}$  leave  $x^0 = 0$  hyperplane invariant and are called *kinematical* while the remaining  $P_0, M_{0k}$  the *dynamical* ones. On the LF there are *seven* kinematical generators:  $P^+, P^1, P^2, B_1, B_2, J_3$  and  $K_3$  which leave the LF hyperplane,  $x^0 + x^3 = 0$ , invariant and the three *dynamical* ones  $S_1, S_2$  and  $P^-$  form a mutually commuting set. The  $K_3$  which was dynamical becomes now a kinematical; it generates scale transformations of the LF components of  $x^\mu, P^\mu$  and  $M^{\mu\nu}$ . We note that each of the set  $\{B_1, B_2, J_3\}$  and  $\{S_1, S_2, J_3\}$  generates an  $E_2 \simeq SO(2) \otimes T_2$  algebra; this will be shown below to be relevant for defining the *spin* for massless particle. Including  $K_3$  in each set we find two subalgebras each with four elements. Some useful identities are  $e^{i\omega K_3} P^\pm e^{-i\omega K_3} = e^{\pm\omega} P^\pm$ ,  $e^{i\omega K_3} P^\perp e^{-i\omega K_3} = P^\perp$ ,  $e^{i\vec{v} \cdot \vec{B}} P^- e^{-i\vec{v} \cdot \vec{B}} = P^- + \vec{v} \cdot \vec{P} + \frac{1}{2}\vec{v}^2 P^+$ ,  $e^{i\vec{v} \cdot \vec{B}} P^+ e^{-i\vec{v} \cdot \vec{B}} = P^+$ ,  $e^{i\vec{v} \cdot \vec{B}} P^\perp e^{-i\vec{v} \cdot \vec{B}} = P^\perp + v^\perp P^+$ ,  $e^{i\vec{u} \cdot \vec{S}} P^+ e^{-i\vec{u} \cdot \vec{S}} = P^+ + \vec{u} \cdot \vec{P} + \frac{1}{2}\vec{u}^2 P^-$ ,  $e^{i\vec{u} \cdot \vec{S}} P^- e^{-i\vec{u} \cdot \vec{S}} = P^-$ ,  $e^{i\vec{u} \cdot \vec{S}} P^\perp e^{-i\vec{u} \cdot \vec{S}} = P^\perp + u^\perp P^-$  where  $P^\perp \equiv \vec{P} = (P^1, P^2)$ ,  $v^\perp \equiv \vec{v} = (v_1, v_2)$  and  $(v^\perp \cdot P^\perp) \equiv (\vec{v} \cdot \vec{P}) = v_1 P^1 + v_2 P^2$  etc. Analogous expressions with  $P^\mu$  replaced by  $X^\mu$  can be obtained if we use  $[P^\mu, X_\nu] \equiv [i\partial^\mu, x_\nu] = i\delta^\mu_\nu$ .

## Appendix B<sup>16</sup>: LF Spin Operator. Hadrons in LF Fock basis

The Casimir generators of the Poincaré group are:  $P^2 \equiv P^\mu P_\mu$  and  $W^2$ , where  $W_\mu = (-1/2)\epsilon_{\lambda\rho\nu\mu}M^{\lambda\rho}P^\nu$  defines the Pauli-Lubanski pseudovector. It follows from  $[W_\mu, W_\nu] = i\epsilon_{\mu\nu\lambda\rho}W^\lambda P^\rho$ ,  $[W_\mu, P_\rho] = 0$  and  $W \cdot P = 0$  that in a representation characterized by particular eigenvalues of the two Casimir operators we may simultaneously diagonalize  $P^\mu$  along with just

<sup>16</sup>See, P.P. Srivastava, *Lightfront quantization of field theory in Topics in Theoretical Physics, Festschrift for Paulo Leal Ferreira*, eds., V.C. Aguilara-Navarro et al., pgs. 206-217, IFT-São Paulo, SP, Brasil (1995); hep-th/9610044; 9610149.

one component of  $W^\mu$ . We have  $W^+ = -[J_3 P^+ + B_1 P^2 - B_2 P^1]$ ,  $W^- = J_3 P^- + S_1 P^2 - S_2 P^1$ ,  $W^1 = K_3 P^2 + B_2 P^- - S_2 P^+$ , and  $W^2 = -[K_3 P^1 + B_1 P^- - S_1 P^+]$  and it shows that  $W^+$  has a special place since it contains only the kinematical generators [15]. On the LF we define  $\mathcal{J}_3 = -W^+/P^+$  as the *spin operator*. It may be shown to commute with  $P_\mu, B_1, B_2, J_3$ , and  $K_3$ . For  $m \neq 0$  we may use the parametrizations  $p^\mu : (p^- = (m^2 + p^\perp{}^2)/(2p^+), p^+ = (m/\sqrt{2})e^\omega, p^1 = -v_1 p^+, p^2 = -v_2 p^+)$  and  $\tilde{p}^\mu : (1, 1, 0, 0)(m/\sqrt{2})$  in the rest frame. We have  $P^2(p) = m^2 I$  and  $W(p)^2 = W(\tilde{p})^2 = -m^2[J_1^2 + J_2^2 + J_3^2] = -m^2 s(s+1)I$  where  $s$  assumes half-integer values. Starting from the rest state  $|\tilde{p}; m, s, \lambda, \dots\rangle$  with  $J_3 |\tilde{p}; m, s, \lambda, \dots\rangle = \lambda |\tilde{p}; m, s, \lambda, \dots\rangle$  we may build an arbitrary eigenstate of  $P^+, P^\perp, \mathcal{J}_3$  (and  $P^-$ ) on the LF by

$$|p^+, p^\perp; m, s, \lambda, \dots\rangle = e^{i(\tilde{v} \cdot \tilde{B})} e^{-i\omega K_3} |\tilde{p}; m, s, \lambda, \dots\rangle$$

If we make use of the following identity [10]

$$\mathcal{J}_3(p) = J_3 + v_1 B_2 - v_2 B_1 = e^{i(\tilde{v} \cdot \tilde{B})} J_3 e^{-i(\tilde{v} \cdot \tilde{B})}$$

we find  $\mathcal{J}_3 |p^+, p^\perp; m, s, \lambda, \dots\rangle = \lambda |p^+, p^\perp; m, s, \lambda, \dots\rangle$ . Introducing also the operators  $\mathcal{J}_a = -(\mathcal{J}_3 P^a + W^a)/\sqrt{P^\mu P_\mu}$ ,  $a = 1, 2$ , which do, however, contain dynamical generators, we verify that  $[\mathcal{J}_i, \mathcal{J}_j] = i\epsilon_{ijk} \mathcal{J}_k$ .

For  $m = 0$  case when  $p^+ \neq 0$  a convenient parametrization is  $p^\mu : (p^- = p^+ v^\perp{}^2/2, p^+, p^1 = -v_1 p^+, p^2 = -v_2 p^+)$  and  $\tilde{p} : (0, p^+, 0^\perp)$ . We have  $W^2(\tilde{p}) = -(S_1^2 + S_2^2)p^{+2}$  and  $[W_1, W_2](\tilde{p}) = 0$ ,  $[W^+, W_1](\tilde{p}) = -ip^+ W_2(\tilde{p})$ ,  $[W^+, W_2](\tilde{p}) = ip^+ W_1(\tilde{p})$  showing that  $W_1, W_2$  and  $W^+$  generate the algebra  $SO(2) \otimes T_2$ . The eigenvalues of  $W^2$  are hence not quantized and they vary continuously. This is contrary to the experience so we impose that the physical states satisfy in addition  $W_{1,2} |\tilde{p}; m = 0, \dots\rangle = 0$ . Hence  $W_\mu = -\lambda P_\mu$  and the invariant parameter  $\lambda$  is taken to define as the *spin* of the massless particle. From  $-W^+(\tilde{p})/\tilde{p}^+ = J_3$  we conclude that  $\lambda$  assumes half-integer values as well. We note that  $W^\mu W_\mu = \lambda^2 P^\mu P_\mu = 0$  and that on the LF the definition of the spin operator appears unified for massless and massive particles. A parallel discussion based on  $p^- \neq 0$  may also be given.

As an illustration consider the three particle state on the LF with the total eigenvalues  $p^+, \lambda$  and  $p^\perp$ . In the *standard frame* with  $p^\perp = 0$  it may be written as  $(|x_1 p^+, k_1^\perp; \lambda_1\rangle |x_2 p^+, k_2^\perp; \lambda_2\rangle |x_3 p^+, k_3^\perp; \lambda_3\rangle)$  with  $\sum_{i=1}^3 x_i = 1$ ,  $\sum_{i=1}^3 k_i^\perp = 0$ , and  $\lambda = \sum_{i=1}^3 \lambda_i$ . Applying  $e^{-i(\tilde{p} \cdot \tilde{B})/p^+}$  on it we obtain  $(|x_1 p^+, k_1^\perp + x_1 p^\perp; \lambda_1\rangle |x_2 p^+, k_2^\perp + x_2 p^\perp; \lambda_2\rangle |x_3 p^+, k_3^\perp + x_3 p^\perp; \lambda_3\rangle)$  now with  $p^\perp \neq 0$ . The  $x_i$  and  $k_i^\perp$  indicate relative (invariant) parameters<sup>17</sup> and do not depend upon the reference frame. The  $x_i$  is the fraction of the total longitudinal momentum carried by the  $i^{th}$  particle while  $k_i^\perp$  its transverse momentum. The state of a pion with momentum  $(p^+, p^\perp)$ , for example, may be expressed as an expansion over the LF Fock states constituted by the different number of partons

$$|\pi : p^+, p^\perp\rangle = \sum_{n, \lambda} \int \bar{\Pi}_i \frac{dx_i d^2 k_i^\perp}{\sqrt{x_i} 16\pi^3} |n : x_i p^+, x_i p^\perp + k_i^\perp, \lambda_i\rangle \psi_{n/\pi}(x_1, k_1^\perp, \lambda_1; x_2, \dots)$$

where [8] the summation is over all the Fock states  $n$  and spin projections  $\lambda_i$ , with  $\bar{\Pi}_i dx_i = \Pi_i dx_i \delta(\sum x_i - 1)$ , and  $\bar{\Pi}_i d^2 k_i^\perp = \Pi_i d^2 k_i^\perp \delta^2(\sum k_i^\perp)$ . The wave function of the parton  $\psi_{n/\pi}(x, k^\perp)$  indicates the probability amplitude for finding inside the pion the partons in the Fock state  $n$  carrying the 3-momenta  $(x_i p^+, x_i p^\perp + k_i^\perp)$ .

The *discrete symmetry* transformations may also be defined on the LF Fock states [8, 15] For example, under the conventional parity  $\mathcal{P}$  the spin operator  $\mathcal{J}_3$  is not invariant. We may rectify this by defining *LF Parity operation* by  $\mathcal{P}^{lf} = e^{-i\pi J_1} \mathcal{P}$ . We find then  $B_1 \rightarrow -B_1, B_2 \rightarrow B_2, P^\pm \rightarrow P^\pm, P^1 \rightarrow -P^1, P^2 \rightarrow P^2$  etc. such that  $\mathcal{P}^{lf} |p^+, p^\perp; m, s, \lambda, \dots\rangle \simeq |p^+, -p^\perp, p^2; m, s, -\lambda, \dots\rangle$ . Similar considerations apply for charge conjugation and time inversion. For example, it is straightforward

<sup>17</sup>We note  $p_i^+ = x_i p^+$ ,  $p_i^\perp = x_i p^\perp + k_i^\perp$ , and  $(p \cdot p) = (2p^+ p^- - p^\perp p^\perp) = \sum_i [(m_i^2 + k_i^\perp k_i^\perp)/x_i]$  where  $(p_i \cdot p_i) = m_i^2$  and  $\sum p_i^\mu = p^\mu$ .

to construct [15] the free *LF Dirac spinor*  $\chi(p) = [\sqrt{2}p^+ \Lambda^+ + (m - \gamma^a p^a) \Lambda^-] \tilde{\chi} / \sqrt{\sqrt{2}p^+ m}$  which is also an eigenstate of  $\mathcal{J}_3$  with eigenvalues  $\pm 1/2$ . Here  $\Lambda^\pm = \gamma^0 \gamma^\pm / \sqrt{2} = \gamma^\mp \gamma^\pm / 2 = (\Lambda^\pm)^\dagger$ ,  $(\Lambda^\pm)^2 = \Lambda^\pm$ , and  $\chi(\tilde{p}) \equiv \tilde{\chi}$  with  $\gamma^0 \tilde{\chi} = \tilde{\chi}$ . The conventional (equal-time) spinor can also be constructed by the procedure analogous to that followed for the LF spinor and it has the well known form  $\chi_{con}(p) = (m + \gamma \cdot p) \tilde{\chi} / \sqrt{2m(p^0 + m)}$ . Under the conventional parity operation  $\mathcal{P} : \chi'(p') = c \gamma^0 \chi(p)$  (since we must require  $\gamma^\mu = L^\mu_\nu S(L) \gamma^\nu S^{-1}(L)$ , etc.). We find  $\chi'(p) = c [\sqrt{2}p^- \Lambda^- + (m - \gamma^a p^a) \Lambda^+] \tilde{\chi} / \sqrt{\sqrt{2}p^- m}$ . For  $p \neq \tilde{p}$  it is not proportional to  $\chi(p)$  in contrast to the result in the case of the usual spinor where  $\gamma^0 \chi_{con}(p^0, -\vec{p}) = \chi_{con}(p)$  for  $E > 0$  (and  $\gamma^0 \eta_{con}(p^0, -\vec{p}) = -\eta_{con}(p)$  for  $E < 0$ ). However, applying parity operator twice we do show  $\chi''(p) = c^2 \chi(p)$  hence leading to the usual result  $c^2 = \pm 1$ . The LF parity operator over spin 1/2 Dirac spinor is  $\mathcal{P}^{lf} = c(2J_1) \gamma^0$  and the corresponding transform of  $\chi$  is shown to be an eigenstate of  $\mathcal{J}_3$ .

### Appendix C: BRS-BFT Quantization on the LF of the CSM

We apply here the recently proposed BFT procedure [60] which is elegant and avoids the computation of Dirac brackets. It would thus get tested [61] on the LF as well and it also allows us to construct (new) effective Lagrangian theories.

We convert the two second class constraints of the bosonized CSM with  $a > 1$  into first class constraints according to the BFT formalism. We obtain then the first class Hamiltonian from the canonical Hamiltonian and recover the DB using Poisson brackets in the extended phase space. The corresponding first class Lagrangian is then found by performing the momentum integrations in the generating functional.

#### (a) Conversion to First Class Constrained Dynamical System

The bosonized CSM model (for  $a > 1$ ) is described by the action

$$S_{CSM} = \int d^2x \left[ -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + \frac{1}{2} \partial_\mu \phi \partial^\mu \phi + e A_\nu (\eta^{\mu\nu} - \epsilon^{\mu\nu}) \partial_\mu \phi + \frac{1}{2} a e^2 A_\mu A^\mu \right], \quad (C. 1)$$

where  $a$  is a regularization ambiguity which enters when we calculate the fermionic determinant in the fermionic CSM. The action in the LF coordinates takes the form

$$S_{CSM} = \int d^2x^- \left[ \frac{1}{2} (\partial_+ A_- - \partial_- A_+)^2 + \partial_- \phi \partial_+ \phi + 2e A_- \partial_+ \phi + a e^2 A_+ A_- \right], \quad (C. 2)$$

We now make the *separation*, in the scalar field (a generalized function) :  $\phi(\tau, x^-) = \omega(\tau) + \varphi(\tau, x^-)$ . The Lagrangian density then becomes

$$\mathcal{L} = \frac{1}{2} (\partial_+ A_- - \partial_- A_+)^2 + \partial_- \varphi \partial_+ \varphi + a e^2 [A_+ + \frac{2}{a e} (\partial_+ \varphi + \partial_+ \omega)] A_-, \quad (C. 3)$$

We note that the dynamical fields are  $A_-$  and  $\varphi$  while  $A_+$  has no kinetic term. On making a redefinition of the (auxiliary) field  $A_+$  we can recast the action on the LF in the following form

$$S_{CSM} = \int dx^- \left[ \dot{\varphi} \varphi' + \frac{1}{2} (\dot{A}_- - A_+' )^2 - 2e \dot{A}_- \varphi + a e^2 A_+ A_- \right], \quad (C. 4)$$

The canonical momenta are given by

$$\begin{aligned} \pi^+ &= 0, \\ \pi^- &= \dot{A}_- - A_+' - 2e\varphi, \\ \pi_\varphi &= \varphi'. \end{aligned} \quad (C. 5)$$



We follow now the Dirac's standard procedure in order to build an Hamiltonian framework on the LF. The definition of the canonical momenta leads to two primary constraints

$$\pi^+ \approx 0, \quad (\text{C. 6})$$

$$\Omega_1 \equiv (\pi_\varphi - \varphi') \approx 0 \quad (\text{C. 7})$$

and we derive one secondary constraint

$$\Omega_2 \equiv \partial_- \pi^- + 2e\varphi' + ae^2 A_- \approx 0. \quad (\text{C. 8})$$

This one follows when we require the  $\tau$  independence (e.g., the persistency) of the primary constraint  $\pi^+$  with respect to the preliminary Hamiltonian

$$H' = H_c^{l.f.} + \int dx u_+ \pi^+ + \int dx u_1 \Omega_1, \quad (\text{C. 9})$$

where  $H_c$  is the canonical Hamiltonian

$$H_c^{l.f.} = \int dx \left[ \frac{1}{2} (\pi^- + 2e\varphi)^2 + (\pi^- + 2e\varphi) A'_+ - ae^2 A_+ A_- \right], \quad (\text{C. 10})$$

and we employ the standard equal- $\tau$  Poisson brackets. The  $u_+$  and  $u_1$  denote the Lagrange multiplier fields. The persistency requirement for  $\Omega_1$  give conditions to determine  $u_1$ . The Hamiltonian is next extended to include also the secondary constraint

$$H_e^{l.f.} = H_c^{l.f.} + \int dx u_+ \pi^+ + \int dx u_1 \Omega_1 + \int dx u_2 \Omega_2 \quad (\text{C. 11})$$

and the procedure is now repeated with respect to the extended Hamiltonian. For the case  $a > 1$ , no more secondary constraints are seen to arise and we are left only with the persistency conditions which determine the multipliers  $u_1$  and  $u_2$  while  $u_+$  is left undetermined. We also find<sup>18</sup>  $\{\Omega_i, \Omega_j\} = D_{ij} (-2\partial_x \delta(x-y))$  where  $i, j = 1, 2$  and  $D_{11} = 1$ ,  $D_{22} = ae^2$ ,  $D_{12} = D_{21} = -e$  and  $\pi^+$  is shown to have vanishing brackets with  $\Omega_{1,2}$ . The  $\pi^+ \approx 0$  constitutes a first class constraint on the phase space; it generates local transformations of  $A_+$  which leave the  $H_e$  invariant,  $\{\pi^+, H_e\} = \Omega_2 \approx 0$ . The  $\Omega_1, \Omega_2$  constitute a set of second class constraints and do not involve  $A_+$  or  $\pi^+$ . It is very convenient, though not necessary, to add to the set of constraints on the phase space the (accessible) gauge fixing constraint  $A_+ \approx 0$ . It is evident from that such a gauge freedom is *not* available at the Lagrangian level. We will also implement (e.g., turn into strong equalities) the (trivial) pair of weak constraints  $A_+ \approx 0$ ,  $\pi^+ \approx 0$  by defining the Dirac brackets with respect to them. It is easy to see that for the other remaining dynamical variables the corresponding Dirac brackets coincide with the standard Poisson brackets. The variables  $A_+, \pi^+$  are thus removed from the discussion, leaving behind a constrained dynamical system with the two second class constraints  $\Omega_1, \Omega_2$  and the light-front Hamiltonian

$$H^{l.f.} = \frac{1}{2} \int dx (\pi^- + 2e\varphi)^2 + \int dx u_1 \Omega_1 + \int dx u_2 \Omega_2 \quad (\text{C. 12})$$

which will be now handled by the BFT procedure.

We introduce the following linear combinations  $\mathbb{T}_i$ ,  $i = 1, 2$ , of the above constraints

$$\begin{aligned} \mathbb{T}_1 &= c_1(\Omega_1 + \frac{1}{M}\Omega_2) \\ \mathbb{T}_2 &= c_2(\Omega_1 - \frac{1}{M}\Omega_2) \end{aligned} \quad (\text{C. 13})$$

<sup>18</sup>We make the convention that the first variable in an equal- $\tau$  bracket refers to the longitudinal coordinate  $x^- \equiv x$  while the second one to  $y^- \equiv y$

where  $c_1 = 1/\sqrt{2(1-e/M)}$ ,  $c_2 = 1/\sqrt{2(1+e/M)}$ ,  $M^2 = ae^2$ , and  $a > 1$ . They satisfy

$$\{\tau_i, \tau_j\} = \delta_{ij}(-2\partial_x \delta(x-y)) \quad (C. 14)$$

and thus diagonalize the constraint algebra.

We now introduce new auxiliary fields  $\Phi^i$  in order to convert the second class constraint  $\tau_i$  into first class ones in the extended phase space. Following BFT [60] we require these fields to satisfy

$$\begin{aligned} \{A^\mu(\text{or } \pi_\mu), \Phi^i\} &= 0, \quad \{\varphi(\text{or } \pi_\varphi), \Phi^i\} = 0, \\ \{\Phi^i(x), \Phi^j(y)\} &= \omega^{ij}(x, y) = -\omega^{ji}(y, x), \end{aligned} \quad (C. 15)$$

where  $\omega^{ij}$  is a constant and antisymmetric matrix. The strongly involutive modified constraints  $\tilde{\tau}_i$  satisfying the abelian algebra

$$\{\tilde{\tau}_i, \tilde{\tau}_j\} = 0 \quad (C. 16)$$

as well as the boundary conditions,  $\tilde{\tau}_i|_{\Phi^i=0} = \tau_i$  are then postulated to take the form of the following expansion

$$\tilde{\tau}_i(A^\mu, \pi_\mu, \varphi, \pi_\varphi; \Phi^j) = \tau_i + \sum_{n=1}^{\infty} \tilde{\tau}_i^{(n)}, \quad \tau_i^{(n)} \sim (\Phi^j)^n. \quad (C. 17)$$

The first order correction terms in this infinite series are written as

$$\tilde{\tau}_i^{(1)}(x) = \int dy X_{ij}(x, y) \Phi^j(y). \quad (C. 18)$$

The first class constraint algebra of  $\tilde{\tau}_i$  then leads to the following condition:

$$\{\tau_i, \tau_j\} + \{\tilde{\tau}_i^{(1)}, \tilde{\tau}_j^{(1)}\} = 0 \quad (C. 19)$$

or

$$(-2\partial_x \delta(x-y))\delta_{ij} + \int dw dz X_{ik}(x, w)\omega^{kl}(w, z)X_{jl}(y, z) = 0. \quad (C. 20)$$

There is clearly some arbitrariness in the appropriate choice of  $\omega^{ij}$  and  $X_{ij}$  which corresponds to the canonical transformation in the extended phase space. We can take without any loss of generality the simple solutions,

$$\begin{aligned} \omega^{ij}(x, y) &= -\delta^{ij}\epsilon(x-y) \\ X_{ij}(x, y) &= \delta_{ij}\partial_x \delta(x-y), \end{aligned} \quad (C. 21)$$

Their inverses are easily shown to be

$$\begin{aligned} \omega^{-1}_{ij}(x, y) &= -\frac{1}{2}\delta_{ij}\partial_x \delta(x-y) \\ (X^{-1})^{ij}(x, y) &= \frac{1}{2}\delta^{ij}\epsilon(x-y), \end{aligned} \quad (C. 22)$$

With the above choice, we find up to the first order

$$\begin{aligned} \tilde{\tau}_i &= \tau_i + \tilde{\tau}_i^{(1)} \\ &= \tau_i + \partial\Phi^i, \end{aligned} \quad (C. 23)$$

and a strongly first class constraint algebra

$$\{\tau_i + \tilde{\tau}_i^{(1)}, \tau_j + \tilde{\tau}_j^{(1)}\} = 0. \quad (C. 24)$$

The higher order correction terms (suppressing the integration operation )

$$\tilde{T}_i^{(n+1)} = -\frac{1}{n+2} \Phi^l \omega^{-1}{}_{lk} (X^{-1})^{kj} B_{ji}^{(n)} \quad (n \geq 1) \quad (C. 25)$$

with

$$B_{ji}^{(n)} \equiv \sum_{m=0}^n \{ \tilde{T}_j^{(n-m)}, \tilde{T}_i^{(m)} \}_{(A, \pi, \varphi, \pi_\varphi)} + \sum_{m=0}^{n-2} \{ \tilde{T}_j^{(n-m)}, \tilde{T}_i^{(m+2)} \}_{(\Phi)} \quad (C. 26)$$

automatically vanish as a consequence of the proper choice of  $\omega^{ij}$  made above. The Poisson brackets are to be computed here using the standard canonical definition for  $A_\mu$  and  $\varphi$  as postulated above. We have now only the first class constraints in the extended phase space and in view of the proper choice only  $\tilde{T}_i^{(1)}$  contributes in the infinite series above.

### (b)- First Class Hamiltonian and Dirac Brackets

We next introduce modified ("gauge invariant") dynamical variables  $\tilde{F} \equiv (\tilde{A}_\mu, \tilde{\pi}^\mu, \tilde{\varphi}, \tilde{\pi}_\varphi)$  corresponding to  $F \equiv (A_\mu, \pi^\mu, \varphi, \pi_\varphi)$  over the phase space by requiring the the following strong involution condition for  $\tilde{F}$  with the first class constraints in our extended phase space, viz,

$$\{ \tilde{T}_i, \tilde{F} \} = 0 \quad (C. 27)$$

with

$$\tilde{F}(A_\mu, \pi^\mu, \varphi, \pi_\varphi; \Phi^j) = F + \sum_{n=1}^{\infty} \tilde{F}^{(n)}, \quad \tilde{F}^{(n)} \sim (\Phi^j)^n \quad (C. 28)$$

and which satisfy the boundary conditions,  $\tilde{F}|_{\Phi^j=0} = F$ .

The first order correction terms are easily shown to be given by

$$\tilde{F}^{(1)}(x) = - \int du dv dz \Phi^j(u) \omega^{-1}{}_{jk}(u, v) X^{-1}{}^{kl}(v, z) \{ T_l(z), F(x) \}_{(A, \pi, \varphi, \pi_\varphi)}. \quad (C. 29)$$

We find

$$\begin{aligned} \tilde{A}_-^{(1)} &= \frac{1}{2M} \partial (c_1 \Phi^1 - c_2 \Phi^2) \\ \tilde{\pi}^{-(1)} &= \frac{M}{2} (c_1 \Phi^1 - c_2 \Phi^2) \\ \tilde{\varphi}^{(1)} &= -\frac{1}{2} (c_1 \Phi^1 + c_2 \Phi^2), \\ \tilde{\pi}_\varphi^{(1)} &= \frac{1}{2} \partial \left[ c_1 \left( 1 - \frac{2e}{M} \right) \Phi^1 + c_2 \left( 1 + \frac{2e}{M} \right) \Phi^2 \right] \end{aligned} \quad (C. 30)$$

where only the combinations  $(c_1 \Phi^1 \pm c_2 \Phi^2)$  of the auxiliary fields are seen to occur. Furthermore, since the modified variables  $\tilde{F} = F + \tilde{F}^{(1)} + \dots$ , up to the first order corrections, are found to be strongly involutive as a consequence of the proper choice made above, the higher order correction terms

$$\tilde{F}^{(n+1)} = -\frac{1}{n+1} \Phi^j \omega_{jk} X^{kl} G_l^{(n)}, \quad (C. 31)$$

with

$$G_l^{(n)} = \sum_{m=0}^n \{ T_i^{(n-m)}, \tilde{F}^{(m)} \}_{(A, \pi, \varphi, \pi_\varphi)} + \sum_{m=0}^{n-2} \{ T_i^{(n-m)}, \tilde{F}^{(m+2)} \}_{(\Phi)} + \{ T_i^{(n+1)}, \tilde{F}^{(1)} \}_{(\Phi)} \quad (C. 32)$$

again vanish. In principle we may follow similar procedure for any functional of the phase space variables; it may get, however, involved.

We make a side remark on the Dirac formulation for dealing with the systems with second class constraints by using the Dirac bracket (DB), rather than extending the phase space. In fact, the Poisson brackets of the modified (gauge invariant) variables  $\tilde{F}$  in the BFT formalism are related [60] to the DB, which implement the constraints  $\mathbb{T}_i \approx 0$  in the problem under discussion, by the relation  $\{f, g\}_D = \{\tilde{f}, \tilde{g}\} |_{\Phi^i=0}$ . In view of only the linear first order correction in CSM the computation of the right hand side is quite simple. We list some of the Dirac brackets

$$\begin{aligned} \{\pi^-, \pi^-\}_D &= \{\widetilde{\pi^-}, \widetilde{\pi^-}\} |_{\Phi=0} \\ &= \{\widetilde{\pi^{-(1)}}, \widetilde{\pi^{-(1)}}\} = \frac{a^2 e^2}{(a-1)} \left(-\frac{1}{4} \epsilon(x-y)\right), \\ \{\varphi, \varphi\}_D &= \{\tilde{\varphi}, \tilde{\varphi}\} |_{\Phi=0} \\ &= \{\tilde{\varphi}^{(1)}, \tilde{\varphi}^{(1)}\} = \frac{a}{(a-1)} \left(-\frac{1}{4} \epsilon(x-y)\right) \\ \{\varphi, \pi^-\}_D &= \{\tilde{\varphi}^{(1)}, \widetilde{\pi^{-(1)}}\} = \frac{ae}{(a-1)} \left(-\frac{1}{4} \epsilon(x-y)\right). \end{aligned} \quad (\text{C. 33})$$

The other ones follow on using the now strong relations  $\Omega_1 = \Omega_2 = 0$  with respect to  $\{, \}_D$  and from  $H_D^{l.f.}$  it follows that the LF Hamiltonian reduces effectively to

$$H_D^{l.f.} = \frac{1}{2} \int dx (\pi^- + 2e\varphi)^2. \quad (\text{C. 34})$$

The first class LF Hamiltonian  $\tilde{H}$  which satisfies the boundary condition  $\tilde{H} |_{\Phi^i=0} = H_D^{l.f.}$  and is in strong involution with the constraints  $\tilde{\mathbb{T}}_i$ , e.g.,  $\{\tilde{\mathbb{T}}_i, \tilde{H}\} = 0$ , may be constructed following the BT procedure or simply guessed for the CSM. It is given by

$$\tilde{H} = \frac{1}{2} \int dx (\tilde{\pi}^- + 2e\tilde{\varphi})^2 \quad (\text{C. 35})$$

which is just the expression in of  $H_D^{l.f.}$  with field variables  $F$  replaced by the  $\tilde{F}$  variables, which already commute with the constraints  $\tilde{\mathbb{T}}_i$ . We do also check that  $\{\tilde{H}, \tilde{H}\} = 0$  and we may identify  $\tilde{H}$  with the BRS Hamiltonian. This completes the operatorial conversion of the original second class system with the Hamiltonian  $H_c$  and constraints  $\Omega_i$  into the first class one with the Hamiltonian  $\tilde{H}$  and (abelian) constraints  $\tilde{\mathbb{T}}_i$ .

### (c)- First Class Lagrangian

We consider now the partition function of the model in order to construct the Lagrangian corresponding to  $\tilde{H}$  in the canonical Hamiltonian formulation discussed above.

We start by representing each of the auxiliary field  $\Phi^i$  by a pair of fields  $\pi^i, \theta^i$ ,  $i = 1, 2$  defined by

$$\Phi^i = \frac{1}{2} \pi^i - \int du \epsilon(x-u) \theta^i(u) \quad (\text{C. 36})$$

such that  $\pi^i, \theta^i$  satisfy

$$\{\pi^i, \theta^j\} = -\delta^{ij} \delta(x-y) \quad \text{etc.}, \quad (\text{C. 37})$$

e.g., the (standard Heisenberg type) canonical Poisson brackets.

Then, The Phase Space Partition Function Is Given By the Faddeev formulae

$$Z = \int \mathcal{D}A_- \mathcal{D}\pi^- \mathcal{D}\varphi \mathcal{D}\pi_\varphi \mathcal{D}\theta^1 \mathcal{D}\pi^1 \mathcal{D}\theta^2 \mathcal{D}\pi^2 \prod_{i,j=1}^2 \delta(\tilde{\mathbb{T}}_i) \delta(\Gamma_j) \det | \{\tilde{\mathbb{T}}_i, \Gamma_j\} | e^{iS}, \quad (\text{C. 38})$$

where

$$S = \int d^2x \left( \pi^- \dot{A}_- + \pi_\varphi \dot{\varphi} + \pi^1 \dot{\theta}^1 + \pi^2 \dot{\theta}^2 - \tilde{\mathcal{H}} \right) \equiv \int d^2x \mathcal{L}, \quad (\text{C. 39})$$

with the Hamiltonian density  $\tilde{\mathcal{H}}$  corresponding to the Hamiltonian  $\tilde{H}$  which is now expressed in terms of  $(\theta^i, \pi_i)$  rather than in terms of  $\Phi^i$ . The gauge-fixing conditions  $\Gamma_i$  are chosen such that the determinants occurring in the functional measure are nonvanishing. Moreover,  $\Gamma_i$  may be taken to be independent of the momenta so that they correspond to the Faddeev-Popov type gauge conditions.

We will now verify in the *unitary gauge*, defined by the original second class constraints:  $\Gamma_i \equiv \Omega_i = 0$ ,  $i=1,2$  being employed in the partition function, do in fact lead to the original Lagrangian. We check that the determinants in the functional measure are non-vanishing and field independent while the product of delta functionals reduces to

$$\delta(\pi_\varphi - \varphi') \delta(\pi^{-'} + 2e\varphi' + M^2 A_-) \delta(\pi^{1'} - 4\theta^1) \delta(\pi^{2'} - 4\theta^2) \quad (\text{C. 40})$$

Since  $\pi_\varphi$  is absent from  $\tilde{H}$  we can perform functional integration over it using the first delta functional. The second delta functional is exponentiated as usual and we name the integration variable as  $A_+$  for convenience. The functional integral over  $\theta^1$  and  $\theta^2$  are easily performed due to the presence of the delta functionals and it also reduces  $\tilde{\mathcal{H}}$  to  $(\pi^- + 2e\varphi)^2/2$ . The functional integrations over the then decoupled variables  $\pi^1$  and  $\pi^2$  give rise to constant factors which are absorbed in the normalization. The partition function in the unitary gauge thus becomes

$$Z = \int \mathcal{D}A_- \mathcal{D}\pi^- \mathcal{D}\varphi \mathcal{D}A_+ e^{iS}, \quad (\text{C. 41})$$

with

$$S = \int d^2x \left[ \pi^- \dot{A}_- + \varphi' \dot{\varphi} + (\pi^{-'} + 2e\varphi' + M^2 A_-) A_+ - \frac{1}{2} (\pi^- + 2e\varphi)^2 \right], \quad (\text{C. 42})$$

Performing the shift  $\pi^- \rightarrow \pi^- - 2e\varphi$  and doing subsequently a Gaussian integral over  $\pi^-$  we obtain the original bosonized Lagrangian with  $\omega$  eliminated by the field redefinition of  $A_+$ . It is interesting to recall that while constructing the LF Hamiltonian framework we eliminated the variable  $A_+$  making use of the gauge freedom on the LF phase space and it gave rise to appreciable simplification. However, on going over to the first class Lagrangian formalism using the partition functional this variable reappears as it should, since the initial bosonized action is not gauge invariant due to the presence of the mass term for the gauge field. Making other acceptable choices for gauge-functions we can arrive at different effective Lagrangians for the system under consideration. It is interesting to recall that in the fermionic Lagrangian the right-handed component of the fermionic field describes a free field and only the left-handed one is gauged. It is also clear from our discussion that  $\tilde{H}$  proposed above is not unique and we could modify it so that it still leads to the original Lagrangian in the unitary gauge. The corresponding first class Lagrangian would produce still other gauge-fixed effective Lagrangians.

## References

- [1] P.A.M. Dirac, Rev. Mod. Phys. 21 (1949) 392.
- [2] We recall the discovery of Kruskal-Szekers coordinates which threw a new light on the problem of the Schwarzschild singularity.  
The  $\pm$  components of a tensor, for example,  $A^\mu$  are defined by  $A^\pm = A_\mp = (A^0 \pm A^3)/\sqrt{2}$  and the metric may be read from  $A \cdot B = A^+ B^- + A^- B^+ - A^1 B^1 - A^2 B^2$ .
- [3] S. Fubini and G. Furlan, Physics. 1 (1964) 229; R. Dashen and M. Gell-Mann, Phys. Rev. Lett. 17 (1966) 340; V. de Alfaro, S. Fubini, G. Furlan, and C. Rossetti, *Currents in Hadron Physics*, North Holland, 1993 and the references cited therein.
- [4] S. Weinberg, Phys. Rev. 150 (1966) 1313.
- [5] J.D. Bjorken, Phys. Rev. 179 (1969) 1547.
- [6] R.P.Feynman, Phys.Rev.Lett. 23 (1969) 1415.
- [7] L. Susskind, Phys. Rev. 165 (1968) 1535; K. Bardakci, M.B. Halpern, *ibid* 176 (1968) 1686; S.J. Chang and S.K. Ma, *ibid* 180 (1969) 1506; H. Leutwyler, Springer Tracts in Mod. Phys. 50 (1969) 29; J. Jersak and J. Stern, Nuovo Cimento 59 (1969) 316; S.D. Drell, J.D. Levy, and T.M. Yan, Phys. Rev. D1 (1970) 1035; F. Rohrlich, Acta Phys. Austr. 32 (1970) 87; J.B. Kogut and D.E. Soper, Phys. Rev. D1 (1970) 2901; J.D. Bjorken, J.B. Kogut, and D.E. Soper, *ibid* D3 (1971) 1382; S.J. Brodsky, R. Roskies, and R. Suaya, *ibid* D8 (1973) 4574; S.J. Chang, R.G. Root and T.M. Yan, Phys. Rev. D7 (1973) 1173; G. 't Hooft, Nucl. Phys. B72 (1974) 461; R. Jackiw, Springer Tracts in Mod. Phys. 62 (1972) 1 and the refs. cited therein.
- [8] S.J. Brodsky, Light-cone quantized QCD and novel hadron phenomenology, SLAC-PUB-7645, 1997; S.J. Brodsky and H.C. Pauli, *Light-cone Quantization and QCD*, Lecture Notes in Physics, vol. 396, eds., H. Mitter et. al., Springer-Verlag, Berlin, 1991; S.J. Brodsky and G.P. Lepage, in *Perturbative Quantum Chromodynamics*, ed., A.H. Mueller, World Scientific, Singapore, 1989;  
C. B. Thorn, Phys. Rev. D20 (1979) 1435; *ibid* Phys. Rev. D20 (1979) 1934.
- [9] K.G. Wilson, T.S. Walhout, A. Harindranath, W.M. Zhang, R.J. Perry, and St. D. Glazek, Phys. Rev. D49 (1994) 6720; K.G. Wilson, Nucl. Phys. B (proc. Suppl.) 17 (1990); R.J. Perry, A. Harindranath, and K.G. Wilson, Phys. Rev. Lett. 65 (1990) 2959.
- [10] P.P. Srivastava, *Light-front Quantization of Field Theory: Some New Results*, Lectures at the IX Brazilian School of Cosmology and Gravitation, July 1998, Rio de Janeiro, Proceedings, Ed. M. Novello, preprint CBPF-NF-003/99, hep-th/9901024; *Lectures on light-front quantized field theory: Spontaneous symmetry breaking. Phase transition in  $(\phi^4)_2$  theory*, Proc. XIV ENPC-Encontro Nacional de Partículas e Campos, Caxambú, MG, pp. 154-192, Sociedade Brasileira de Física, São Paulo, SP, Brasil, 1993; hep-th/ 9312064; Nuovo Cimento A107 (1994) 549. See [8, 9, 10, 32] for the extensive list of references.
- [11] R.J. Perry, *Light-front quantum chromodynamics*, nucl-th/9901080; Hadrons '94, Eds. V. Herscovitz et al., World Scientific, Singapore, 1995.
- [12] See [32, 10].
- [13] D. Bigatti and L. Susskind, *Review of matrix theory*, hep-th/9712072; Phys. Lett. B425 (1998) 351, hep-th/9711063.
- [14] E. Witten, Commun. Math. Phys. 92 (1984) 455.
- [15] P.P. Srivastava, Mod. Phys. Letts. A13 (1998) 1223; See also, P.P. Srivastava, in *Geometry, Topology and Physics*, Apanasov et. al. (Eds.), Walter de Gruyter & Co., Berlin, New York, 1997, pp. 260; hep-th/9610149 and 9610044.

- [16] P.P. Srivastava, Phys. Letts. B448 (1999) 68; hep-th9811225.
- [17] P.P. Srivastava and S.J. Brodsky, *Light-front quantized QCD in covariant gauge*, SLAC-PUB-8168; hep-ph/9906423.
- [18] P.P. Srivastava, Nuovo Cimento A108 (1995) 35; see also [10].
- [19] A.N. Mitra and S. Bhatnagar, Int. J. Mod. Phys. A7 (1992) 121.
- [20] A.N. Mitra, Phys. Lett. B (1999); hep-ph/9812404, and hep-ph/9901421;
- [21] J. Carbonel et al., Phys. Rep. 400 (1998) 215.
- [22] See, S.S. Schweber, *Relativistic Quantum Field Theory*, Row, Peterson and Co., New York, 1961; J.D. Bjorken and S.D. Drell, *Relativistic Quantum Fields*, McGraw Hill, 1965; L.H. Ryder, *Quantum Field Theory*, Cambridge University Press, 2nd Edition, 1996.
- [23] On the physical grounds we must require the cluster decomposition principle, which requires that distant experiments give uncorrelated results. See, S. Weinberg, in *Conceptual foundations of quantum field theory*, Ed. T.Y. Cao, Cambridge University Press, 1999; N.N. Bogolubov et.al., *Introduction to Axiomatic quantum field theory*, Benjamin, 1975.  
The *locality* does not seem to be strictly required; the *front form* theory may show nonlocality (Sec. 1.1) along the longitudinal direction even when the corresponding *instant form* theory is formulated as a local theory.
- [24] See for example, I.N. Sneddon, *Elements of Partial Differential Equations*, McGraw-Hill, NY, 1957, pp. 111-115
- [25] P.A.M. Dirac, *Lectures in Quantum Mechanics*, Belfer Graduate School of Science, Yeshiva University Press, New York, 1964; Can. J. Math. 2, 129 (1950); E.C.G. Sudarshan and N. Mukunda, *Classical Dynamics: a modern perspective*, Wiley, NY, 1974. See also L. Faddeev and R. Jackiw, Phys. Rev. Lett. 60 (1988) 1692.
- [26] S. Coleman, Commun. Math. Phys. 31 (1973) 259.
- [27] The LF components of the four-momentum are  $k^\mu = (k^-, k^+, k^\perp)$  where  $k^\pm = (k^0 \pm k^3)/\sqrt{2}$ . Here  $k^-$  is the LF energy while  $k^\perp$  and  $k^+$  indicate the transverse and the *longitudinal* components of the momentum respectively. For a free massive particle on the mass shell we have the dispersion relation:  $2k^-k^+ = (k^\perp{}^2 + m^2) > 0$  so that  $k^\pm$  are both positive when  $k^0 > 0$  or both negative when  $k^0 < 0$ . It has no square root as found in  $k^0 = \pm\sqrt{k^\perp{}^2 + m^2}$ . The conservation of the total longitudinal momentum does not permit the excitation of massive quanta by the LF vacuum which has vanishing longitudinal momentum. It should, however, be noted that when dealing with the momentum space loop integrals, a significant contribution may arise from such configuration in the integrand; the reason being that we have to deal with the products of several distributions. The components  $(k^1, k^2, k^3)$  in the *instant form* theory on the other hand may take positive or negative values and the conventional theory vacuum state may contain an arbitrary number of particles (and antiparticles) which may mix with the vacuum state, with no particles, to form the ground state.
- [28] P.P. Srivastava and E.C.G. Sudarshan, Phys. Rev. 110 (1958) 765.
- [29] E. Fermi, Prog. Theo. Phys. (Japan) 5 (1950) 570; Phys. Rev. 92 (1953) 452.
- [30] For massless particles the correlation ceases to exist at the point  $p^\perp \rightarrow 0$  since  $2p^+p^- = p^\perp p^\perp \rightarrow 0$ .
- [31] J. Barcelos Neto and C. Wotzasek, Europhys. Lett. 21 (1993) 511; R. Amorim and J. Barcelos Neto, Zeit. Phys. C68 (1995) 513; F.P. Devecchi and H.O. Girotti, Phys. Rev. D49 (1994) 4302 and refs. contained therein.

- [32] Such constraints on the potential, as illustrated by Dirac [1] in his paper, are required when we unify in relativistic theory the principles of special relativity and the principles of quantization. It is interesting to note that soon after in 1950-52 he formulated also the systematic method (Dirac procedure) for constructing Hamiltonian formulation for constrained dynamical system. That the constraint (3.1) resulted in *nonlocal LF Hamiltonian* and consequently gave a description of SSB were pointed out in : P.P. Srivastava, *On spontaneous symmetry breaking mechanism on the light-front quantized field theory*, Ohio-State University preprint 91-0481, Slac database no. PPF-9148, November 1991 and the references cited therein. The broken continuous symmetry in  $3 + 1$  dimensions and the *tree level Higgs mechanism* is discussed in Ohio-State University preprint 92-0012, Slac database no. PPF-9202, December, 1991. *Spontaneous symmetry breaking mechanism on the light-front quantized field theory- Discretized formulation*, Ohio-State University preprint 92-0173, Slac database no. PPF-9222, April 1992, available as scanned copies on the Spires hep-th data base. See also the papers contributed to *XXVI Intl. Conference on High energy Physics, Dallas, Texas*, August 1992, *AIP Conf. Proc.*, 272 (1993) 2125, Ed. J.R. Sanford, database: csum c92/08/06, conf(uspires-slac), papers 135, 136; University of Padova, Report No. DFPF/92/TH/58.
- [33] H.C. Pauli and S.J. Brodsky, *Phys. Rev. D* 32 (1985) 2001.
- [34] P.P. Srivastava, *Light-front quantization and Spontaneous Symmetry Breaking- Discretized formulation, Hadron Physics 94*, pp. 253, Eds. V. Herscovitz et. al., World Scientific, Singapore, 1995; hep-th/9412204, 205.
- [35] G. Parisi, *Statistical Field Theory*, Addison-Wesley, 1988.
- [36] B. Simon and R.B. Griffiths, *Commun. Math. Phys.* 33 (1973) 145.
- [37] See, J. Leite Lopes, *Gauge Field Theories*, Pergamon Press, 1981.
- [38] N. Marcus and J. Schwarz, *Phys. Lett.* 115B (1982) 111; D.J. Gross, J.A. Harvey, E. Martinec, and R. Rohm, *Phys. Rev. Lett.* 54 (1985) 502.
- [39] X.G. Wen, *Phys. Rev. Lett.* 64 (1990) 2206; M. Stone, *Phys. Rev.* B41 (1990) 212.
- [40] W. Siegel, *Nucl. Phys.* B238 (1984) 307.
- [41] R. Floreanini and R. Jackiw, *Phys. Rev. Lett.* 59 (1987) 1873.
- [42] P.P. Srivastava, *Phys. Rev. Lett.* 63 (1989) 2791.
- [43] P.P. Srivastava, *in preparation*.
- [44] M.E. Costa and H.O. Girotti, *Phys. Rev. Lett.* 60 (1988) 1771.
- [45] A. Tseytlin and P. West, *Phys. Rev. Lett.* 65 (1990) 541.
- [46] See Barcellos et al. in [31].
- [47] W.T. Kim, J.K. Kim, and Y.J. Park, *Phys. Rev. D* 44 (1991) 563.
- [48] C. Imbimbo and A. Schwimmer, *Phys. Lett.* B193 (1987) 455; J.M.F. Labastida and M. Pernici, *Phys. Rev. Lett.* 59 (1987) 2511.
- [49] H.O. Girotti, M. Gomes, and V.O. Rivelles, *Phys. Rev. D* 45 (1992) R3329; D.S. Kulshreshtha and H.J.W. Muller-Kirsten, *Phys. Rev. D* 45 (1992) R393.
- [50] In the conventional metric,  $\eta^{\mu\nu} = \text{diag} (1, -1, -1, -1)$ ,  $\mu, \nu = 0, 1, 2, 3$ , the  $\gamma$  matrices are defined as usual,  $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}$ ,  $\gamma^0 \gamma^\mu \gamma^0 = \gamma^{\mu\dagger}$ ,  $C \gamma^\mu C^{-1} = -\gamma^{\mu T}$ ,  $C = i\gamma^2 \gamma^0$ ,  $\Sigma_3 = \Sigma_3^\dagger = i\gamma^1 \gamma^2$ ,  $\Sigma_2 = i\gamma^3 \gamma^1$ ,  $\Sigma_1 = i\gamma^2 \gamma^3$ ,  $\gamma_5 = i\gamma^0 \gamma^1 \gamma^2 \gamma^3 = \gamma_5^\dagger$ ,  $\gamma_5^2 = I$ ,  $[\gamma_5, \vec{\Sigma}] = 0$  etc. *No explicit representation of  $\gamma^\mu$  matrices is used in our discussions.*



- [51] Here it is understood that an unsymmetrical expression like  $\bar{\psi}\gamma^\mu\partial_\mu\psi$  is to be replaced by its symmetrized form  $[\bar{\psi}\gamma^\mu\partial_\mu\psi - \partial_\mu\bar{\psi}\gamma^\mu\psi]/2$ . It is convenient to work on the LF in terms of the projected four spinors  $\psi_\pm$ .
- [52] J. Tiomno, *Il Nuovo Cimento*, 1 (1956) 226.
- [53] See R. Jackiw in ref. [7]
- [54] P.P. Srivastava, *Europhys. Lett.* 33 (1996) 423 and *LF dynamics of Chern-Simons systems*, ICTP, Trieste preprint IC/94/305; hep-th/9412239.
- [55] See, A. Dhar, G. Mandal and S.R. Wadia, *Phys. Lett.* B329 (1994) 15; D.J. Gross and I. Klebanov, *Nucl. Phys.* B352 (1990) 671.
- [56] See, S.J. Brodsky et al. [8] and references cited therein.
- [57] See for example, E. Abdalla, M.C. Abdalla and K. Rothe, *Non-Perturbative Methods in Two Dimensional Quantum Field Theory*, World Scientific, Singapore, 1991 and the references cited therein; D. Boyanovsky, *Nucl. Phys.* B294 (1987) 223; A. Bassetto, L. Griguolo, and P. Zanca, *Phys. Rev.* D50 (1994) 1077.
- [58] R. Jackiw, R. Rajaraman, *Phys. Rev. Lett.* 54 (1985) 1219, 2060(E).
- [59] A. Bassetto, G. Nardelli, and R. Soldati, *Yang-Mills theories in algebraic noncovariant gauges*, World Scientific, 1991; G. Leibbrandt, *Noncovariant gauges, Quantization of Yang-Mills and Chern-Simons theory in axial-type gauges*, World Scientific, 1994.
- [60] I.A. Batalin and I.V. Tyutin, *Int. J. Mod. Phys.* A6 (1991) 3255.
- [61] P.P. Srivastava, *BRS-BFT quantization of the CSM on the light-front*, paper LP-002, Session P17, *Intl. Symp. on Lepton-Photon Interactions- LP'97*, July 1997, Hamburg. Available as .ps file on the DESY database or as hep-th/9901024.

# 18. Gauge Symmetry In Chiral Electrodynamics

D.S.Kulshreshtha \*

Department of Physics & Astrophysics.  
University of Delhi, Delhi-110007, India

## Abstract

The constrained dynamics and local vector gauge invariance of the various field theory models describing the chiral electrodynamics in one-space one-time dimension is revisited.

## 1. Introduction

The one-space one-time ((1+1)-) dimensional field theories [1-26] are the simplest toy models which are exactly soluble and renormalizable and their study has given rise to a very important concept namely, that any fermion (plus boson if desired) field theory always has its boson equivalent field theory called the bosonized field theory [3]. As a consequence of this equivalence called bosonization (or fermionization) many interesting features of two-dimensional (2D) field theories have been revealed [1-26]. This concept of bosonization (originally discovered in the context of 2D field theories) has also been very useful in the understanding of four-dimensional phenomena that may be described by an effective 2D theory [27]. The rather convincing demonstration of quark confinement in exactly soluble 2D models is perhaps one of the best known examples of the successes of this field [3]. Another outstanding idea discovered in this field is that the vector gauge boson of chiral electrodynamics (CED) has an anomaly generated mass [7,8], in contrast to the picture of spontaneous symmetry breaking with Higgs mechanism for the vector boson mass generation in the standard model. The 2D theories could be physically relevant in many circumstances. The 2D conformally invariant quantum field theories e.g., describe the long range behaviour of correlations in planar statistical systems undergoing second order phase transitions [28]. There are also some physical systems whose motion is dynamically constrained to lie in a subspace of the full space-time and a lower dimensional model adequately describes the reduced dynamics. Linear polymers like polyacetylene e.g., have been described by the use of an anomalous 2D field theory [29]. The so-called anomalous field theories are infact, the theories that do not possess the gauge symmetry [30] so that they lack the gauge invariance. We would take up these ideas in details in the next section. The last (but not least) reason is that these 2D field theory models are the basic building blocks of some larger theories in the field of string theories [29]. The plan of the article is as follows. In Secs. 2 and 3, we discuss some basics about the concepts related to the gauge symmetry and CED. In Secs. 4, 5 and 6, we consider some specific theories namely, the chiral Schwinger models (CSM's) describing 2D-CED. In Sec. 7, we briefly discuss the self-dual fields called the chiral bosons (CB's) which represent the chiral Fermi theories, and we finally summarise the article in the last section.

## 2. The Gauge Symmetry

The existence of symmetries in physics lead to rather important conservation laws. The so-called symmetries are infact, the transformations that leave the action of the theory unchanged or invariant [30]. One way of classifying the symmetries that exist in nature is in terms of the so-called

---

\*Email : dsk@physics.du.ac.in

external or space-time symmetries and the internal symmetries. External symmetries include space- and -time translational invariance (leading to the conservation of linear momentum and energy) as well as the invariance under Lorentz transformations including rotations and the boosts (leading to the conservation of angular momentum and Lorentz-boosts). In contrast to the above examples, the example of gauge invariance which leads to the existence of conserved currents and conserved charges, is an example of the so-called internal symmetries [30]. In the present article, we would consider the implications of the continuous (and in particular, the internal) symmetries in the framework of Lagrangian field theories. For this purpose, let us consider field theory defined by the action integral functional :

$$S = \int d^D x \mathcal{L}(\phi_k, \partial_\mu \phi_k); \mu = 0, 1, \dots, (D-1) \quad (1)$$

where  $\mathcal{L} \equiv \mathcal{L}(\phi_k, \partial_\mu \phi_k)$  is the Lagrangian density (LD) of the theory and D is the dimension of the space-time in the Minkowski space. We now consider the transformation of the coordinates :

$$x_\mu \longrightarrow x'_\mu = x_\mu + \delta x_\mu \quad (2)$$

The space-time point  $x'$  is in general a function of  $x$  and  $\delta x_\mu (= x'_\mu - x_\mu)$ , includes infinitesimal translations as well as rotations. This transformation induces a transformation in the fields  $\phi_k(x)$  :

$$\phi_k(x) \longrightarrow \phi'_k(x') = \phi_k(x) + \hat{\delta}\phi_k(x) \quad (3)$$

where  $\hat{\delta}\phi_k(x)$  is a small symmetry transformation. For the case  $x' = x$ , the transformation is called an internal transformation. In this case the space-time point remains unchanged so that one has  $\delta x_\mu = 0$  for an internal transformation. In this case ( $x'_\mu = x_\mu$ ) one has :

$$\hat{\delta}\phi_k(x) = \delta\phi_k(x) = \phi'_k(x) - \phi_k(x) = [i\beta^a T_a \phi_k(x)]; a = 1, 2, \dots, N \quad (4)$$

where  $T_a$  are the generators of the symmetry transformation ( $\delta\phi(x)$ ) of  $\phi(x)$  and  $\beta^a$  are the N gauge parameters corresponding to N independent transformations. Further, if  $\beta^a = \beta^a(x_\mu)$  is an arbitrary function of  $x_\mu$  then the symmetry transformation is called a local transformation, where as  $\beta^a = \text{constant}$  leads to the so-called global symmetry transformations. Also with any symmetry transformation there exists in general, an associated Noether current  $J_\mu^a(x)$ , which is conserved if the symmetry is exact so that  $\delta S = 0$ , giving rise to the continuity equation :

$$\partial^\mu J_\mu^a(x) = 0 \quad (5)$$

for the conserved Noether current  $J_\mu^a(x)$  (which is called the vector gauge current (VGC) when the symmetry is a vector gauge symmetry (VGS)) leading in turn to the existence of the conserved (global) charge  $Q^a(t)$  :

$$Q^a = Q^a(t) = \int d^{D-1}x J_0^a(x) \quad (6)$$

One of the important experimental tests of a theory is whether the conserved quantities it predicts are indeed conserved. Infact, the identification of the transformations of the fields that leave the action of the theory invariant leads to important predictions of the theory without solving the equations of motion. The proper intuition about (5) can be obtained from the usual ED where the electromagnetic current satisfies a continuity equation which says that the charge is neither created nor destroyed locally [30]. Equation (5) only generalizes this result of ED to other kind of charges. A crucial feature of field theories with LGI is that for each independent internal LVGS there exists a vector gauge field and its corresponding vector boson particle that mediates (or carries the force) between the charged matter fields. For ED the gauge field is the electromagnetic vector potential  $A_\mu(x)$  and its quantum particle is the massless spin-1 photon. In fact, all interactions in general are mediated by vector bosons originating from the local symmetries which dictate the exact form of interaction. The interaction Lagrangian must be of the form  $[e J^\mu(x) A_\mu(x)]$ , where the coupling constant  $e$  is defined as the strength with which the vector boson  $A_\mu(x)$  interacts with the VGC  $J^\mu(x)$  [30]. Maxwell's unification of electricity and magnetism can, infact, be viewed as the discovery that ED is described by the simplest possible LVGS implying the corresponding invariance. Maxwell's addition of the displacement current to the field equations (the Maxwell's equations) which was made to ensure the conservation of electromagnetic current turns out to be equivalent to imposing LGI on the Lagrangian of ED [30].

Although the above discussions are classical, the results are usually correct in the quantum theory derived from a classical Lagrangian. In some cases, however the quantum corrections contribute a non-zero term to the right hand side of the continuity equations (5) and these terms are called anomalies. For global symmetries these anomalies can often improve the predictions from Lagrangians that have too much symmetries when compared with data because anomalies wreck the symmetry that was never present in the quantum theory even though the classical Lagrangian had the symmetry. However, for local symmetries (LVGS) presence of anomalies is rather disastrous [30,31]. A quantum theory is locally symmetric only if its gauge currents satisfy the continuity equation (5), otherwise the local anomalies simply change the theory. In view of this one has to deal rather carefully with this kind of gauge anomalous field theories [30-31]. The main object of the present article is to investigate some of the gauge anomalous theories in (1+1)-dimension describing the CED [1-26] and to show as to how one could try to restore the LVGS or the LGI to the otherwise gauge anomalous theory. An understanding of the constrained dynamics of a system is usually found to be very helpful in this context. The LVGS of a dynamical system, as we would see in the following, is very intimately connected with the constrained dynamics of the dynamical system [31], and the Hamiltonian formulation à la Dirac [32] is, particularly suited to discover the dynamical generators of internal symmetries of a constrained dynamical system. The gauge-invariant (GI) systems in general play an important role in the theoretical description of the fundamental laws of nature. In fact, most of the physical systems of interest e.g., the ED, quantum ED (QED), QCD, electro-weak theory and the gravity theory, are all constrained systems [30-31]. A constrained dynamical system, in fact, is one which is defined in terms of the over determined set of coordinates and the Hamiltonian formulation makes it easier to keep track of all the coordinates, canonical and redundant where the complete set of constraints emerges naturally. It may be important to mention here that the nature of the matrix of the poisson brackets (PB's) of the constraints of the theory, as we would see in the later sections, determines the nature of the set of constraints of the theory and also as to whether the theory is GI or not. Thus if the above matrix is singular, then the set of constraints of the theory is first-class and the theory is GI (and also if this matrix is a null matrix (and therefore also singular) then the theory is a true or bonafide GI theory). On the other hand, if this matrix is non-singular then the set of constraints of the theory is second-class and the theory is gauge-non-invariant (GNI). This, in fact, could even be taken as a criterion for differentiating the GI systems from the GNI ones [4,5,9,10,13,14,18,19,21,22,25,26]. These GI systems could then be quantized under some appropriate gauge choices or the gauge-fixing conditions (GFC's) [30-31]. Further, in the usual Hamiltonian formulation of a GI theory under some GFC's, one necessarily destroys the gauge invariance of the theory by fixing the gauge (which converts a set of first-class constraints into a set of second-class constraints, implying a breaking of LGI under the gauge-fixing). To achieve the quantization of a GI theory such that the gauge invariance of the theory is maintained even under gauge-fixing, one goes to a more generalized procedure called the Becchi-Rouet-Stora and Tyutin (BRST) formulation [33]. In the BRST formulation of a GI theory, the theory is rewritten as a quantum system that possesses a generalized gauge invariance called the BRST symmetry [33]. For this, one enlarges the Hilbert space of the GI theory and replaces the notion of the gauge transformation, which shifts operators by c-number functions, by a BRST transformation, which mixes operators having different statistics. In view of this, one introduces new anti-commuting variables called the Faddeev-Popov ghost and anti-ghost fields, which are Grassmann numbers on the classical level and operators in the quantized theory, and a commuting variable called the Nakanishi-Lautrup field. In the BRST formulation, one thus embeds a GI theory into a BRST invariant system, and quantum Hamiltonian of the system (which includes the gauge-fixing contribution) commutes with the BRST charge operator as well as with the anti-BRST charge operator and the new symmetry of the quantum system (the BRST symmetry) that replaces the gauge invariance is maintained (even under the gauge-fixing) and hence projecting any state onto the sector of BRST and anti-BRST invariant states yields a theory which is isomorphic to the original GI theory [31,33]. The unitarity and consistency of the BRST-invariant theory described by the gauge-fixed quantum Lagrangian is guaranteed by the conservation and nilpotency of the BRST charge.

Also, the relativistic quantum dynamics of a physical system could be studied either in the conventional formulation on the hyperplane  $x^0 = \text{constant}$ , called the instant-form (IF) [34] or on the hyperplanes of the light-front :  $(x^0 + x^1) = \text{constant}$ , called the front-form (FF) à la Dirac [34]. In the present work, we would study the theories describing CED in both the forms of dynamics.

### 3. Chiral Electrodynamics

The Schwinger model describing ED in one-space one-time dimension with massless fermions and its chiral versions called the chiral Schwinger models (CSM's) have been of a very wide interest in the recent years [1–26]. The CSM describes a massless Dirac field  $\psi(x, t)$  in two dimensions with only one of its chiral components coupled to a U(1) gauge field  $A^\mu(x, t)$  [7,10]. The first CSM was introduced by Hagen [6] as a new example of an exactly solvable field theory in the (1+1)-dimension. In this model the gauge field with a bare mass was considered [6]. Jackiw and Rajaraman [7,8], later on, considered the gauge anomalous theory without the bare mass term. By studying the field equations and propagator obtained from the effective gauge field action, they concluded [7,8] that the theory was not gauge invariant, but was unitary and amenable to particle interpretation [7,8]. They also found that the vector gauge boson necessarily acquires a mass when consistency and unitarity are demanded [7,8]. One of the remarkable achievements of the studies of such theories as mentioned at the beginning has been the development of the fermion-boson correspondence in 2D quantum field theories. The other important achievement has been in the field of understanding the phenomena of gauge anomalies and the gauge anomalous field theories [7–22]. The JR-CSM [7,8] is found to admit exact solutions in a positive metric Hilbert space, respecting unitarity, provided the JR regularization parameter  $a$  (introduced in Ref. [7]) is restricted to the range  $a \geq 1$  [7,8], for which the theory is sensible. Infact, the model is seen to yield a sensible theory for a class of regularizations [7,10]. The spectrum of the theory depends on the regularization in a crucial way and it is seen to contain, for  $a > 1$ , a massive photon in addition to a massless fermion, and for  $a = 1$ , only a massless fermion [7,8]. The JR-CSM is seen to lack the LVGS and is therefore gauge anomalous. It is rather well known that corresponding to a GNI theory, a GI theory could be constructed by the inclusion of the so-called Stueckelberg term (ST)/Wess-Zumino term (WZT) [35] in the action of the GNI theory. The JR-CSM has been studied rather widely in the recent years [7–10]. In particular, the Hamiltonian and BRST formulations of the GI versions of this theory have been studied in the IF in Refs. [8,9], and in the FF in Ref. [10], where the GI versions of this theory have also been constructed by the inclusion of an appropriate ST/WZT. The physical contents of the original GNI theory are also recovered under a special choice of gauge and the equivalence of the quantized GI and GNI theories is established.

Very recently, Mitra [11] has considered a new regularization [11,12] which does not belong to the above class. With this regularization [11,12] the photon is once again massive and the massless fermion present in the theory has (unlike the JR-regularization) a chirality opposite to that entering the interaction with the electromagnetic field [11,12]. Further, this regularization, being in accordance with the Faddeev's picture [35] of anomalous gauge theories, has been called by Mitra [11] as the Faddeevian regularization [11–14].

If the matrix of the PB's of the constraints of the theory becomes non-singular because of the non-vanishing PB of the Gauss law constraint of the theory with itself (called Faddeev's anomaly [35]), so that the constraints become second-class and the theory becomes GNI (or it loses LGI) because of this Faddeev's anomaly, then the theory fits into the Faddeev's scenario [35]. In the CSM with the Faddeevian regularization considered by Mitra [11,12], the Faddeevian mechanism works because the constraints of the theory become second class through the Faddeev's anomaly for the Gauss law constraint of the theory. The above CSM with the Faddeevian regularization has been studied in the IF in Refs. [11,12,14], where the IF theory is seen to be GNI possessing a set of three second-class constraints. The corresponding FF theory is also found to be GNI [13] possessing a set of three second-class constraints. The Hamiltonian and BRST formulations of the GI versions of this theory have been studied in the IF in Ref. [14], and in the FF in Ref. [13]. The Mitra-CSM with the Faddeevian regularization has a mass-like term for the vector gauge boson

$A_\mu$  different than those of the class of models called JR-CSM's and may be taken as a signature of new regularization [11-14]. This theory in contrast with the JR-CSM is seen to possess a self-dual boson which could also be thought of as a chiral fermion.

Yet another important CSM is due to Harada [15], who has, in particular, constructed a gauged Floreanini-Jackiw [16] action which describes a CSM [15-19] in terms of chiral bosonization (as explained below) [15]. This model is seen to possess a set of three second-class constraints and consequently it describes a GNI or a gauge anomalous theory. An appropriate ST/WZT for this GNI-theory in the IF has been calculated in Ref. [18] where the Hamiltonian and BRST formulations of the resulting GI (and consequently a gauge non-anomalous theory) have also been studied. Now while solving the minimal CSM (the CSM where the right-handed (or equivalently the left-handed) fermion is absent), one is faced with a technical difficulty. In order to construct an operator solution, one usually makes use of the knowledge of a solution of the bosonized model. Because the minimal CSM contains only the left-handed fermion, one is not able to obtain its bosonized form by the usual bosonization, and one is led to consider its chiral bosonization [15]. Harada [15] has considered the chiral bosonization of the minimal CSM and has obtained a gauged action [15] corresponding to the Floreanini-Jackiw action [16], from the conventional bosonic one of the CSM, by imposing the chiral constraint ( $\pi - \phi \approx 0$  (cf. Ref. [15])) in phase space. After obtaining this equivalent bosonic action, the equations of motion have been solved in the GNI formulation. The bosonic solution is found to be completely satisfactory. Harada has further constructed [15] a fermionic operator solution of the minimal CSM in covariant gauges in the GI formulation [15]. He found a free chiral fermion (self dual chiral boson) and a free massive scalar (boson) with the desired mass, as physical asymptotic fields in a positive-definite Hilbert space. The existence of a physical free chiral fermion distinguishes the CSM from the (vector) Schwinger model [1-5] (which has no physical asymptotic fermion) and implies that the fermion is not confined. This minimal CSM is also found to be completely consistent like the usual JR-CSM [6-10]. The Harada's CSM in the FF [19] is seen to possess a set of three first-class constraints and consequently describes a GI theory, in contrast to the IF theory [15,18] which is GNI owing to the second-class nature of the set of constraints of the theory. The Hamiltonian and BRST formulations of this FF theory have been studied in Ref. [19]. In section 6, we would consider this theory in the IF as well as in the FF, in details.

#### 4. The Generalized Schwinger Model

The generalized Schwinger model (GSM) which describes the QED in (1+1)-dimension with massless fermions is described by the LD [1-5,7-10] :

$$\mathcal{L}_f = [i\bar{\psi}\gamma^\mu\partial_\mu\psi + \frac{1}{2}e_R\bar{\psi}\gamma^\mu(1+\gamma^5)\psi A_\mu + \frac{1}{2}e_L\bar{\psi}(1-\gamma^5)\psi A_\mu - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}] \quad (7a)$$

$$g_1 = \frac{1}{2}(e_L - e_R) ; \quad g_2 = \frac{1}{2}(e_L + e_R) ; \quad (7b)$$

$$\gamma^5 = \gamma^0\gamma^1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \gamma^\mu\gamma^5 = -\varepsilon^{\mu\nu}\gamma_\nu \quad (7c)$$

$$g^{\mu\nu} := \text{diag}(+1, -1) ; \quad \varepsilon^{\mu\nu} = -\varepsilon^{\nu\mu} ; \quad \varepsilon^{01} = +1 ; \quad \mu, \nu = 0, 1 \quad (7d)$$

which is equivalent to its bosonized form [1-5,7-10] :

$$\mathcal{L}_b = [\frac{1}{2}\partial_\mu\phi\partial^\mu\phi + (g_1g^{\mu\nu} - g_2\varepsilon^{\mu\nu})\partial_\mu\phi A_\nu - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{M^2}{2\pi}A_\mu A^\mu] \quad (8)$$

In the rest of the article we would work with the above conventions. The mass term for  $A_\mu$  here arises from the regularization ambiguities associated with the definition of the current. The case of CSM is obtained from the GSM by setting  $g_1 = g_2 = g$  (i.e.  $e_R = 0$ ) ; and  $M^2 = g^2 a$ , where  $a$  is the JR-regularization parameter. The case of vector Schwinger model (VSM) is obtained by setting  $g_1 = 0, g_2 = g$  (i.e.  $e_L = e_R$ ) ; and  $M = 0$ . Here  $e_L = e_R$  implies a vector-like theory in this case [1-5]. Also in the case of VSM, demanding the regularization to be gauge-invariant fixes  $a = 0$  i.e.  $M = 0$ . On the other hand, in CSM [7-10], no choice of  $a$  can make the theory GI and therefore  $a$  is left as a free parameter. The Hamiltonian and BRST formulations of the vector [1-5] and chiral [7-10] theories in the IF have been studied in Refs. [4,9] and in the FF in Refs. [5,10]. The VSM in the bosonized form is described by the LD [1-5] :

$$\mathcal{L} := [\frac{1}{2}\partial_\mu\phi\partial^\mu\phi - g\varepsilon^{\mu\nu}\partial_\mu\phi A_\nu - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}] ; \quad \mu, \nu = 0, 1 \quad (9)$$

Equation (11) describes the theory in the IF [1,4] and is seen to possess a set of two constraints :  $\Omega_1 = \Pi_o \approx 0$  and  $\Omega_2 = (E + g\phi) \approx 0$ . Where  $\Omega_1$  is a primary constraint and  $\Omega_2$  is a secondary constraint. Here,  $\Pi_o$ ,  $E (= \Pi^1)$  and  $\Pi$  are the momenta canonically conjugate respectively to  $A_o$ ,  $A_1$  and  $\phi$ . The divergence of the VGC for the IF theory is seen to vanish implying that the theory possess (at the classical level) a VGS. The matrix of the PB's of these constraints is seen to be singular implying that the set of the constraints is first-class and that the theory is GI. The action of the theory is infact, seen to be invariant under the LVGT :  $\delta\phi = 0$ ,  $\delta A_o = \beta$ ,  $\delta A_1 = \beta$ ,  $\delta\Pi = -\beta$ ,  $\delta\Pi_o = \delta E = 0$ . Here  $\beta = \beta(x, t)$  is an arbitrary function of the coordinates. The theory could thus be quantized under some gauge choice. The Hamiltonian and BRST formulations of this theory in the IF have been studied in Ref. [4]. In the FF [34] one defines the light-cone coordinates :  $x^\pm := (x^0 \pm x^1) / \sqrt{2}$ , and then writes all the quantities involved in the LD in terms of  $x^\pm$  instead of  $x^0$  and  $x^1$ . After doing this the LD in the FF reads [5] :

$$\mathcal{L} = [(\partial_+ \phi)(\partial_- \phi) + g(\partial_+ \phi)A^+ - g(\partial_- \phi)A^- + \frac{1}{2}(\partial_+ A^+ - \partial_- A^-)^2] \quad (10a)$$

$$A^\mp = A_\pm = (A_o \pm A_1) / \sqrt{2} \text{ and } \partial_\pm \phi = (\dot{\phi} \pm \phi) / \sqrt{2} \quad (10b)$$

The VGC is seen to be conserved, i.e.,  $\partial_\mu J^\mu := [\partial_+ J_- + \partial_- J_+] = 0$ , for the above theory, implying that the theory possesses (at the classical level) a VGS. The theory is seen to possess a set of three constraints:  $\chi_1 = (\Pi^+) \approx 0$ ;  $\chi_2 = (\Pi - \partial_- \phi - gA^+) \approx 0$ ;  $\chi_3 = (\partial_- \Pi - g(\partial_- \phi)) \approx 0$ . Where  $\chi_1$  and  $\chi_2$  are the primary constraints and  $\chi_3$  is a secondary constraint. Here  $\Pi^+$ ,  $\Pi^-$  and  $\Pi$  are the momenta canonically conjugate respectively to  $A^-$ ,  $A^+$  and  $\phi$ . The matrix of the PB's of the constraints  $\chi_i$  is first-class and that the theory described by  $\mathcal{L}(12)$  is GI. The action of the theory is, in fact, seen to be invariant under the LVGT [5] :  $\delta A^+ = \partial_- \beta$ ,  $\delta A^- = \partial_+ \beta$ ,  $\delta\phi = 0$ ,  $\delta u = \partial_+ \partial_+ \beta$ ,  $\delta v = 0$ ;  $\delta\Pi^+ = 0$ ,  $\delta\Pi^- = 0$ ,  $\delta\Pi = g\partial_- \beta$ ;  $\delta\Pi_u = 0$ ,  $\delta\Pi_v = 0$ . Where  $\beta = \beta(x^-, x^+)$  is an arbitrary function of the coordinates. The theory could therefore be quantized under some appropriate gauge choice.

The JR-CSM in the IF is described in the bosonized form by the LD (with  $\mu, \nu = 0, 1$ ) [7-9] :

$$\mathcal{L} = [\frac{1}{2}\partial_\mu \phi \partial^\mu \phi + e(g^{\mu\nu} - \epsilon^{\mu\nu})\partial_\mu \phi A_\nu - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{1}{2}ae^2 A_\mu A^\mu] \quad (11)$$

In (13), the first term corresponds to a massless boson, which is equivalent to a massless fermion. The second term represents the chiral coupling of this fermion to the electromagnetic field  $A_\mu$ . The third term is the kinetic energy term and the last term is the mass term for the vector gauge boson and contains the signature of regularization where  $a$  is the JR-regularization parameter. In the following we would set the coupling constant  $e = 1$ . For the case  $a = 1$ , the above theory is seen to possess a set of four second-class constraints [7-9] :  $\Omega_1 = \Pi_o \approx 0$ ;  $\Omega_2 = (E + \phi + \Pi + A_1) \approx 0$ ;  $\Omega_3 = E \approx 0$ ;  $\Omega_4 = (-\Pi - \phi - 2A_1 + A_o) \approx 0$ . Where  $\Omega_1$  is a primary constraint and  $\Omega_2, \Omega_3, \Omega_4$ , are the secondary constraints. For the case  $a \neq 1$ , however, the theory possesses a set of only two second-class constraints :  $\zeta_1 = \Pi_o \approx 0$ ;  $\zeta_2 = [E + \phi + \Pi + A_1 + (a-1)A_o] \approx 0$ . As a consequence of the second-class nature of the constraints  $\Omega_i$  (for the case  $a = 1$ ) and of the constraints  $\zeta_i$  (for the case  $a \neq 1$ ), the model describes in the IF of dynamics, for both the cases  $a = 1$  and  $a \neq 1$ , a GNI theory [7-9]. For constructing a GI theory corresponding to the GNI model  $\mathcal{L}^N$ , we calculate the appropriate ST/WZT [35]  $\mathcal{L}^S$  for the theory, the addition of which to  $\mathcal{L}^N$  restores the GI to the theory. The ST/WZT, for the JR-CSM, e.g., for the case  $a = 1$ , is obtained by enlarging the Hilbert space of the GNI theory, by introducing a new field  $\theta$  called the Stueckelberg/Wess-Zumino field through the following redefinition of fields  $\phi$  and  $A^\mu$  in  $\mathcal{L}^N$  :  $\phi \rightarrow (\phi - \theta)$  and  $A^\mu \rightarrow (A^\mu + \partial^\mu \theta)$ . The ST/WZT thus obtained is [7-9] :

$$\mathcal{L}^S = (\theta A_1 + \theta A_o). \quad (12)$$

The new GI theory so obtained is defined by the LD :  $\mathcal{L}^I = (\mathcal{L}^N + \mathcal{L}^S)$  and is seen to possess a set of four first-class constraints :  $\psi_1 = \Pi_o \approx 0$ ,  $\psi_2 = (\Pi_\theta - A_1) \approx 0$ ,  $\psi_3 = (E + \phi + \Pi + A_1 - \theta) \approx 0$  and  $\psi_4 = E \approx 0$ . The matrix of PB's of  $\psi_i$  is seen to be singular implying that the theory is GI. The theory is indeed invariant under the LVGT [9] :  $\delta\phi = -\beta$ ,  $\delta A_o = \beta$ ,  $\delta A_1 = \beta$ ,  $\delta\theta = -\beta$ ,  $\delta\Pi = -\beta$ ,  $\delta\Pi_\theta = \beta$ ,  $\delta E = \delta\Pi_o = 0$ . The theory could thus be quantized under some appropriate gauge choice. However, in order to recover the physical contents of the original GNI theory, we go to a special gauge  $\partial^\mu \theta = 0$  (or equivalently,  $\theta = 0$  and  $-\theta = 0$ ), and accordingly we choose the GFC's :  $\zeta_1 = -\theta = 0$  and  $\zeta_2 = (-\Pi - \phi - 2A_1 + A_o) \approx 0$ . It is easy to see that  $\mathcal{L}^I$  under the above gauge, reproduces precisely the quantum system described by  $\mathcal{L}^N$ . So that this gauge translates the GI version of the theory into the GNI one. Infact, the physical Hilbert spaces of the



$A_\mu$  different than those of the class of models called JR-CSM's and may be taken as a signature of new regularization [11-14]. This theory in contrast with the JR-CSM is seen to possess a self-dual boson which could also be thought of as a chiral fermion.

Yet another important CSM is due to Harada [15], who has, in particular, constructed a gauged Floreanini-Jackiw [16] action which describes a CSM [15-19] in terms of chiral bosonization (as explained below) [15]. This model is seen to possess a set of three second-class constraints and consequently it describes a GNI or a gauge anomalous theory. An appropriate ST/WZT for this GNI-theory in the IF has been calculated in Ref. [18] where the Hamiltonian and BRST formulations of the resulting GI (and consequently a gauge non-anomalous theory) have also been studied. Now while solving the minimal CSM (the CSM where the right-handed (or equivalently the left-handed) fermion is absent), one is faced with a technical difficulty. In order to construct an operator solution, one usually makes use of the knowledge of a solution of the bosonized model. Because the minimal CSM contains only the left-handed fermion, one is not able to obtain its bosonized form by the usual bosonization, and one is led to consider its chiral bosonization [15]. Harada [15] has considered the chiral bosonization of the minimal CSM and has obtained a gauged action [15] corresponding to the Floreanini-Jackiw action [16], from the conventional bosonic one of the CSM, by imposing the chiral constraint ( $\pi - \phi \approx 0$  (cf. Ref. [15])) in phase space. After obtaining this equivalent bosonic action, the equations of motion have been solved in the GNI formulation. The bosonic solution is found to be completely satisfactory. Harada has further constructed [15] a fermionic operator solution of the minimal CSM in covariant gauges in the GI formulation [15]. He found a free chiral fermion (self dual chiral boson) and a free massive scalar (boson) with the desired mass, as physical asymptotic fields in a positive-definite Hilbert space. The existence of a physical free chiral fermion distinguishes the CSM from the (vector) Schwinger model [1-5] (which has no physical asymptotic fermion) and implies that the fermion is not confined. This minimal CSM is also found to be completely consistent like the usual JR-CSM [6-10]. The Harada's CSM in the FF [19] is seen to possess a set of three first-class constraints and consequently describes a GI theory, in contrast to the IF theory [15,18] which is GNI owing to the second-class nature of the set of constraints of the theory. The Hamiltonian and BRST formulations of this FF theory have been studied in Ref. [19]. In section 6, we would consider this theory in the IF as well as in the FF, in details.

#### 4. The Generalized Schwinger Model

The generalized Schwinger model (GSM) which describes the QED in (1+1)-dimension with massless fermions is described by the LD [1-5,7-10] :

$$\mathcal{L}_f = [i\bar{\psi}\gamma^\mu\partial_\mu\psi + \frac{1}{2}e_R\bar{\psi}\gamma^\mu(1+\gamma^5)\psi A_\mu + \frac{1}{2}e_L\bar{\psi}(1-\gamma^5)\psi A_\mu - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}] \quad (7a)$$

$$g_1 = \frac{1}{2}(e_L - e_R) ; \quad g_2 = \frac{1}{2}(e_L + e_R) ; \quad (7b)$$

$$\gamma^5 = \gamma^0\gamma^1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \gamma^\mu\gamma^5 = -\varepsilon^{\mu\nu}\gamma_\nu \quad (7c)$$

$$g^{\mu\nu} := \text{diag}(+1, -1) ; \quad \varepsilon^{\mu\nu} = -\varepsilon^{\nu\mu} ; \quad \varepsilon^{01} = +1 ; \quad \mu, \nu = 0, 1 \quad (7d)$$

which is equivalent to its bosonized form [1-5,7-10] :

$$\mathcal{L}_b = [\frac{1}{2}\partial_\mu\phi\partial^\mu\phi + (g_1g^{\mu\nu} - g_2\varepsilon^{\mu\nu})\partial_\mu\phi A_\nu - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{M^2}{2\pi}A_\mu A^\mu] \quad (8)$$

In the rest of the article we would work with the above conventions. The mass term for  $A_\mu$  here arises from the regularization ambiguities associated with the definition of the current. The case of CSM is obtained from the GSM by setting  $g_1 = g_2 = g$  (i.e.  $e_R = 0$ ) ; and  $M^2 = g^2 a$ , where  $a$  is the JR-regularization parameter. The case of vector Schwinger model (VSM) is obtained by setting  $g_1 = 0, g_2 = g$  (i.e.  $e_L = e_R$ ) ; and  $M = 0$ . Here  $e_L = e_R$  implies a vector-like theory in this case [1-5]. Also in the case of VSM, demanding the regularization to be gauge-invariant fixes  $a = 0$  i.e.  $M = 0$ . On the other hand, in CSM [7-10], no choice of  $a$  can make the theory GI and therefore  $a$  is left as a free parameter. The Hamiltonian and BRST formulations of the vector [1-5] and chiral [7-10] theories in the IF have been studied in Refs. [4,9] and in the FF in Refs. [5,10]. The VSM in the bosonized form is described by the LD [1-5] :

$$\mathcal{L} := [\frac{1}{2}\partial_\mu\phi\partial^\mu\phi - g\varepsilon^{\mu\nu}\partial_\mu\phi A_\nu - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}] ; \quad \mu, \nu = 0, 1 \quad (9)$$



$\delta u = \partial_o \partial_o \beta$ ,  $\delta v = -\partial_o \beta$ ;  $\delta \Pi = \delta \Pi_o = \delta E = \delta \Pi_\theta = \delta \Pi_u = \delta \Pi_v = 0$ . Where  $\beta = \beta(x, t)$  is an arbitrary function of the coordinates. This GI theory under the gauge:  $\zeta_1 = -\partial_1 \theta \approx 0$ ;  $\zeta_2 = [(A_o + A_1) - (\Pi + \Pi_\theta) + e(\Pi + \partial_1 \phi)] \approx 0$ , reproduces precisely the quantum system described by  $\mathcal{L}^N$  (14). The Hamiltonian and BRST quantizations of this GI theory have been studied in the IF in Ref. [14]. Further, the original theory  $\mathcal{L}^N$  (14) in the FF is described by the LD [13]:

$$\mathcal{L}^N := [\partial_+ \phi](\partial_- \phi) + 2eA^+(\partial_+ \phi) + \frac{1}{2}(\partial_+ A^+ - \partial_- A^-)^2 - e^2(A^-)^2 + 2e^2 A^+ A^-];$$

$$\mu, \nu = 0, 1. \quad (16)$$

The above theory is seen to possess a set of three second-class constraints  $\rho_i$ :  $\rho_1 = (\Pi^+) \approx 0$ ;  $\rho_2 = (\Pi - \partial_- \phi - 2eA^+) \approx 0$ ;

$\rho_3 = [\partial_- \Pi^- + 2e^2(A^- - A^+)] \approx 0$ , and is therefore GNI. Corresponding to this GNI theory a GI theory could be constructed by the inclusion of an appropriate ST/WZT [13]:

$$\mathcal{L}^S = [(1 - 2e + 2e^2)(\partial_+ \theta)(\partial_- \theta) - (1 - 2e)(\partial_+ \phi)(\partial_- \theta) - (\partial_+ \theta)(\partial_- \phi) + 2e(e - 1)A^+(\partial_+ \theta) - e^2(\partial_+ \theta)^2 - 2e^2 A^- (\partial_+ \theta - \partial_- \theta)] \quad (17)$$

The resulting GI theory is seen to possess at the classical level, a LVGS, and it is also seen to possess a set of three first-class constraints:  $\psi_1 := \Pi^+ \approx 0$ ;  $\psi_2 := [\Pi - \partial_- \phi - 2eA^+ + (1 - 2e)\partial_- \theta] \approx 0$ ; and  $\psi_3 = [\partial_- \Pi^- + \Pi_\theta - (1 - 2e)(\partial_- \theta) + \partial_- \phi + 2eA^+] \approx 0$ , where  $\psi_1$  and  $\psi_2$  are the primary constraints and  $\psi_3$  is the secondary constraint implying that the theory is GI. It is indeed seen to be invariant under the LVGT [13]:  $\delta A^+ = \partial_- \beta$ ,  $\delta A^- = \partial_+ \beta$ ,  $\delta \phi = -\beta$ ,  $\delta \theta = -\beta$ ,  $\delta u = \partial_+ \partial_+ \beta$ ,  $\delta v = -\partial_+ \beta$ ;  $\delta \Pi^+ = \delta \Pi^- = \delta \Pi = \delta \Pi_\theta = \delta \Pi_u = \delta \Pi_v = 0$ . Where  $\beta = \beta(x^+, x^-)$  is an arbitrary function of its arguments. This GI theory under the gauge [13]:  $\zeta_1 = -(\partial_- \theta) \approx 0$ ;  $\zeta_2 = [\Pi_\theta + \partial_- \phi - 2e(e - 1)A^+ + 2e^2 A^-] \approx 0$ , reproduces precisely the quantum GNI system described by (16). The Hamiltonian and BRST formulations of this theory have been investigated in Ref. [13] under some specific gauge choices, where this GI theory has been constructed through the ST/WZT given by (17).

## 6. The CSM Due to Harada

In this section we consider a CSM in terms of chiral bosonization constructed by Harada [15] by gauging the Floreanini-Jackiw action [16] and described in the IF by the LD [15,18]:

$$\mathcal{L}^N = [(\partial_o \phi - \partial_1 \phi)\partial_1 \phi + 2e(A_o - A_1)(\partial_1 \phi) - \frac{1}{2}e^2(A_o - A_1)^2 - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{1}{2}ae^2 A_\mu A^\mu]; \quad \mu, \nu = 0, 1. \quad (18)$$

This theory is seen to possess a set of three second-class constraints:  $\Omega_1 = \Pi_o \approx 0$ ;  $\Omega_2 = (\Pi - \phi) \approx 0$ ;  $\Omega_3 = (E + 2e\phi + e^2\{(a - 1)A_o + A_1\}) \approx 0$ . Where  $\Omega_1$  and  $\Omega_2$  are the primary constraints and  $\Omega_3$  is a secondary constraint. The above set of constraints is second-class and consequently the theory is GNI. A GI-theory corresponding to this GNI theory is obtained by the addition of an appropriate ST/WZT [18]:

$$\mathcal{L}^S = [(\partial \theta - \partial \phi - \partial \phi) + (2\phi\theta - \theta^2) + 2e(\partial - \theta)(\phi - \theta) - 2e(A_o - A_1)\theta - \frac{1}{2}e^2(\partial - \theta)^2 - e^2(A_o - A_1)(\partial - \phi) + \frac{1}{2}ae^2(\theta^2 + 2A_o\theta - 2A_1\theta - \theta^2)] \quad (19)$$

The resulting GI-theory is seen to possess a set of three first-class constraints [18]:  $\psi_1 = \Pi_o \approx 0$ ;  $\psi_2 = [\Pi - \phi + \theta] \approx 0$ ;  $\psi_3 = [E + \phi + \Pi_\theta - \theta] \approx 0$ . The matrix of PB's of  $\psi_i$  is seen to be singular implying that the theory is GI and it is indeed seen to be invariant under the LVGT [18]:  $\delta \phi = -\beta$ ,  $\delta A_o = \beta$ ,  $\delta A_1 = \beta$ ,  $\delta \theta = -\beta$ ,  $\delta \Pi = \delta \Pi_o = \delta E = \delta \Pi_\theta = 0$ . Further, this GI theory under the gauge:  $\zeta_1 = -\theta = 0$ ;  $\zeta_2 = [\Pi_o - (2e - 1)\phi + e^2(A_o - A_1) - ae^2 A_o] \approx 0$ ; reproduces precisely the physical contents of the original quantum GNI theory described by (18) [18]. The Hamiltonian and BRST formulations of the GI version of the Harada's CSM obtained by the inclusion of the appropriate ST/WZT have been studied in the IF in Ref. [18].

The Harada's CSM in the FF reads [19]:

$$\mathcal{L} = [(\partial_- \phi)(\partial_+ \phi - \partial_- \phi) + 2eA^+(\partial_+ \phi - \partial_- \phi) - e^2(A^+)^2 + \frac{1}{2}(\partial_+ A^+ - \partial_- A^-)^2 + ae^2 A^+ A^-] \quad (20)$$

The VGC for (20) is seen to be conserved, and the theory is seen to possess (at the classical level) a LVGS [19]. Also, it is seen to possess a set of three first-class constraints:  $\chi_1 := (\Pi^+) \approx 0$ ;  $\chi_2 := (\Pi - \partial_- \phi - 2eA^+) \approx 0$ ;  $\chi_3 := (\partial_- \Pi^- + ae^2 A^+) \approx 0$ . Where  $\chi_1$  and  $\chi_2$  are the

primary constraints and  $\chi_3$  is a secondary constraint. The matrix of the PB's of the constraints  $\chi_i$  is singular implying that the set of constraints  $\chi_i$  is first-class and that the theory is GI. The theory is indeed seen to be invariant under the LVGT [19]:  $\delta\phi = -e\beta$ ,  $\delta A^+ = \partial_- \beta$ ,  $\delta A^- = \partial_+ \beta$ ,  $\delta u = \partial_+ \partial_+ \beta$ ,  $\delta v = -e\partial_+ \beta$ ;  $\delta\Pi = e\partial_- \beta$ ,  $\delta\Pi^- = \delta\Pi^+ = \delta\Pi_u = \delta\Pi_v = 0$ . Where  $\beta = \beta(x^+, x^-)$  is an arbitrary function of its arguments. The Hamiltonian and BRST formulations of this theory have been studied under some specific gauge choices in Ref. [19].

## 7. The Chiral Bosons

Towards the end let us discuss in brief two examples of the self-dual fields called chiral bosons which represent the chiral Fermi theories. These fields are basic building blocks of the larger theories in the CED and the string theories. The first example that we consider is that of the Srivastava chiral bosons [20], which are single self-dual fields and are described by the LD [20-22] :

$$\mathcal{L}^N = [\frac{1}{2}\partial_\mu\phi\partial^\mu\phi + \lambda_\mu(\epsilon^{\mu\nu} + g^{\mu\nu})\partial_\nu\phi] \quad (21)$$

The above theory is seen to possess a set of two second-class constraints :  $\Omega_1 = p_\lambda \approx 0$  ; and  $\Omega_2 = (\Pi - \phi' - \lambda) \approx 0$  ; where  $\lambda = (\lambda_0 + \lambda_1)$  ; and consequently the theory is GNI. Corresponding to this GNI theory a GI theory could be obtained by the addition of an appropriate ST/WZT [21]:

$$\mathcal{L}^S = [-\frac{1}{2}\dot{\phi}^2 - \frac{1}{2}\phi'^2 + \phi\theta' + \dot{\theta}\phi - \dot{\theta}\phi' - \lambda(\dot{\theta} - \theta)] ; \quad (22)$$

to  $\mathcal{L}^N$ . The resulting GI-theory is seen to possess a set of two first-class constraints :  $\psi_1 = p_\lambda \approx 0$  ; and  $\psi_2 = (\Pi + \Pi_0) \approx 0$  ; and it is indeed seen to be invariant under the LVGT [21] :  $\delta\phi = \pm\beta$ ,  $\delta\theta = \pm\beta$ ,  $\delta\lambda = \mp\beta$ ,  $\delta\Pi = \delta\Pi_0 = \delta p_\lambda \approx 0$ . Now this GI theory under the gauge :  $\zeta_1 = -\theta' \approx 0$  ;  $\zeta_2 = (\Pi_0 - \theta' + \phi' + \lambda) \approx 0$ ; reproduces precisely the original GNI quantum system. The other example is that of the so-called Siegel chiral bosons which are doubly self-dual and are described by the LD [23-26] :

$$\mathcal{L} = [\frac{1}{2}\dot{\phi}^2 - \frac{1}{2}\phi'^2 + \lambda(\dot{\phi} - \phi')^2] \quad (23)$$

This theory is seen to possess one primary constraint  $\Omega_1 = p_\lambda \approx 0$  , and one secondary constraint  $\tilde{\Omega}_2 = \frac{1}{(1+2\lambda)^2}(\Pi - \phi')^2 \approx 0$  which is classically equivalent to  $\Omega_2 = (\Pi - \phi') \approx 0$ . The matrix of the PB's of  $\Omega_1$  and  $\Omega_2$  is singular, implying that the set of constraints  $\Omega_1$  and  $\Omega_2$  is first-class and that the theory is GI. The theory is indeed seen to be invariant under the so-called Siegel gauge transformations (which are LVGT) [23-26] :  $\delta\phi = \beta(\dot{\phi} - \phi')$ ;  $\delta p_\lambda = 0$ ;  $\delta\lambda = [-\frac{1}{2}(\dot{\beta} + \beta) + \beta(\dot{\lambda} - \lambda) - \lambda(\dot{\beta} - \beta)]$ ;  $\delta\Pi = \beta[(1+2\lambda)(\dot{\phi} - \phi') - 2\lambda(\dot{\phi} - \phi') + 2(\dot{\phi} - \phi')(\dot{\lambda} - \lambda)] - \beta[(\dot{\phi} - \phi)']$ . Where the gauge parameter  $\beta = \beta(x, t)$  is an arbitrary function of its arguments, and  $\lambda$  can be made equal to any given function of  $x$  and  $t$ , for an appropriate choice of  $\beta$ . The Hamiltonian formulation of the theory has been studied in Ref.. [25] and its BRST formulation in Ref. [26], under some specific gauge choices.

## 8. Summary and Discussions

In this article, we have studied the concept of the LVGS as interlinked with the constrained dynamics of the various field theory models describing the 2D-CED. The constrained dynamics of some specific CSM's describing the CED have been considered in the IF as well as in the FF of dynamics in details. The chiral bosons representing the chiral Fermi theories have also been briefly considered towards the end.

The author thanks Dr. Usha Kulshreshtha for her long time collaboration on this work and for several useful discussions.

## References

- [1] J.Schwinger, Phys. Rev., **128**, 2425 (1962).
- [2] See e.g., K.R.Ito, Prog. Theor. Phys., **53** , 817 (1975); N.Nakanishi, Prog.Theor. Phys., **57** , 580 (1977); K.D.Rothe and J.A.Swieca, Phys. Rev., **D15**, 541 (1977); J.H.Lowenstein and

- A.Swieca, *Ann. Phys.*, **68**, 172 (1971); S.Coleman, R.Jackiw and L.Susskind, *Ann. Phys.*, **93**, 267 (1975); S.Coleman, *Ann. Phys.*, **101**, 239 (1976); C.Wotzasek, *Acta Phys. Pol.*, **B21**, 457 (1990); J.McCabe, *Phys. Lett.*, **257B**, 145 (1991).
- [3] A.Casher, J.Kogut and L.Susskind, *Phys. Rev. Lett.*, **31**, 792(1973); *Phys.Rev.*, **D10**, 732 (1974); J.Kogut and L.Susskind,*Phys. Rev.*, **D11**, 3594 (1975); *ibid*, **D13**, 337 (1976).
- [4] Usha Kulshreshtha,D.S.Kulshreshtha and H.J.Mueller-Kirsten, *Helv. Phys. Acta*, **66**, 737 (1993); Usha Kulshreshtha,D.S.Kulshreshtha and H.J.W. Mueller-Kirsten, "Hamiltonian and BRST formulation of some chiral Field Theories in one-space one-time Dimension", Proceedings of II workshop on "Const- raints Theory and Quantization Methods" Montepulciano (Siena) Italy, 1993, World Scientific Singapore, 305-327 (1994).
- [5] Usha Kulshreshtha and D.S.Kulshreshtha, *Intl. J.Theor. Phys.*, **37**, 2539 (1998).
- [6] C.Hagen, *Ann. Phys.*, (N.Y.), **81**, 65 (1973).
- [7] R.Jackiw and R.Rajaraman, *Phys. Rev. Lett.*, **54**, 1219 (1985); **54**, 2060 (E) (1985); R.Rajaraman, *Phys. Lett.*, **154B**, 305 (1985); *ibid* **184B**,369 (1987).
- [8] H.O.Girotti, H.J.Rothe and K.D.Rothe, *Phys. Rev.*, **D33**, 514 (1986); N.K.Falck, G.Kramer, *Ann. Phys.*, **176**, 330 (1987); P.Mitra and R.Rajaraman, *Phys. Lett.*, **225B**, 267 (1989); *Ann. Phys.*,**203**, 157 (1990); D.Boyanovski, *Nucl. Phys.*, **B294**, 223(1987); R.P.Malik,*Phys. Lett.*, **212B**, 445 (1988); P.P.Srivastava, *Phys. Lett.*,**235B**, 287 (1990); D.Boyanovsky, I.Schmidt and M.F.L.Golterman, *Ann. Phys. (N.Y.)*, **185**, 111(1988).
- [9] Usha Kulshreshtha, D.S.Kulshreshtha and H.J.Mueller-Kirsten, *Can. J.Phys.*, **72**, 639 (1994); *Il. Nuovo Cim.*,**A107**, 569 (1994).
- [10] Usha Kulshreshtha and D.S.Kulshreshtha, *Canad. J. Phys.*, **77**, (1999)-in press.
- [11] . P.Mitra, *Phys. Lett.*, **284B**, 23 (1992).
- [12] S.Mukhopadhyay and P.Mitra, *Zeit. f. Phys.***C97**, 525 (1995); S.Mukhopadhyay and P.Mitra, *Ann. Phys. (N.Y.)* **241**, 68 (1995).
- [13] Usha Kulshreshtha, *Helv. Phys. Acta* **71**, 353 (1998).
- [14] Usha Kulshreshtha, "A Gauge Invariant Chiral Schwinger Model with the Faddeevian Regularization: The Stueckelberg Term, Hamiltonian and BRST Formulations", Delhi University, Preprint-1999.
- [15] K. Harada, *Phys. Rev. Lett.*, **64**, 139 (1990); *Phys. Rev.*, **D42**, 4170 (1990).
- [16] R.Floresani, R.Jackiw, *Phys. Rev. Lett.*, **59**, 1873 (1987).
- [17] J.K.Kim,W.T.Kim,W.H.Kye, *Phys. Rev.*, **D42**, 4170 (1990); *Phys. Rev.* **D45**, 717 (E) (1992); W.H.Kye, W.T.Kim, J.K.Kim, *Phys. Lett.*, **B268**, 59 (1991).
- [18] Usha Kulshreshtha, D.S.Kulshreshtha and H.J.Mueller-Kirsten, *Zeit. f. Phys.*, **C64**, 169 (1994).
- [19] Usha Kulshreshtha, *Canad. J. Phys.* **78**, (2000)-in press.
- [20] P.P.Srivastava, *Phys. Rev. Lett.*, **63**, 2791 (1989); *Phys. Lett.*, **234B**, 93 (1990).
- [21] Usha Kulshreshtha, D.S.Kulshreshtha and H.J.Mueller-Kirsten, *Zeit. f. Phys.*, **C60**, 427 (1993).
- [22] Usha Kulshreshtha, D.S.Kulshreshtha and H.J.Mueller-Kirsten, *Phys. Rev.*, **D45**, R393 (1992).

- [23] W.Siegel, Nucl. Phys., **B238**, 307 (1984).
- [24] J.M.F.Labastida and M.Pernichi, Phys. Rev. Lett., **59**, 2511 (1987); Nucl. Phys. **B297**, 557 (1988); S.Bellucci, M.F.L.Goltermann and D.N. Petcher, Nucl. Phys. **B326**, 307 (1989).
- [25] Usha Kulshreshtha, D.S.Kulshreshtha and H.J.Mueller-Kirsten, Phys. Rev., **D47**, 4634 (1993).
- [26] Usha Kulshreshtha and D.S.Kulshreshtha, Intl. J. Theor. Phys., **38**,1399 (1999).
- [27] V.Rubakov, Nucl. Phys., **B203**, 311(1982); C.G.Callan, Phys. Rev., **D25**, 2141 (1982).
- [28] J.L.Cardy, "Phase Transitions and Critical Phenomena", Vol.11, Academic Press, London (1985).
- [29] M.B.Green, J.H.Schwarz and E.Witten, "Superstring Theory", Vol.1 & 2,Cambridge University Press, Cambridge (1987).
- [30] C.Itzykson and J.Zuber, "Quantum Field Theory", Intl. Series in Pure and Applied Physics, McGraw-Hill (1980).
- [31] M.Henneaux and C.Teitelboin, "Quantization of Gauge Systems", Prince ton University Press, Princeton, New Jersey (1992).
- [32] P.A.M.Dirac, Can. J. Math., **2**, 129 (1950); "Lectures on Quantum Mechanics", Belfer Graduate School of Science, Yashiva University, New York, 1964.
- [33] C.Becchi, A.Rouet and R.Stora, Phys. Lett., **52B**, 344 (1974) ; V.Tyutin Lebedev Report No. FIAN-39 (1975), D.Nameschansky, C.Preitschopf and M.Weinstein, Ann. Phys., (N.Y.), **183**, 226 (1988) ;D.M.Gitman, "Quantization of Fields with Constraints", Springer-Verlag (1990).
- [34] P.A.M.Dirac, Rev. of Mod. Phys., **21**, 392 (1949).
- [35] E.C.G.Stueckelberg, Helv. Phys. Acta, **14**, 52 (1941); Helv. Phys. Acta **30**, 209 (1957); J.Wess and B.Zumino, Phys. Lett.,**37B**, 95 (1971); E.Witten, Nucl. Phys., **B223**, 422 (1983); L.D.Faddeev,Phys. Lett., **145B**, 81 (1984); L.D.Faddeev and S.L. Shatashvili, Phys. Lett., **167B**, 225 (1986); P.Mitra and R.Rajaraman, Phys. Lett., **225B**, 267 (1989); Ann. Phys. (N.Y.),**203**, 157 (1990).

# 19. Towards a Unified Description of the Four Interactions in Terms of Dirac-Bergmann Observables.

Luca Lusanna \*

[3mm]Sezione INFN di Firenze,  
Largo E.Fermi 2, 50125 Firenze, Italy

## Abstract

A review is given of the status and developments of the research program aiming to reformulate the physics of the four interactions at the classical level in a unified way in terms of Dirac-Bergmann observables with special emphasis on the open mathematical, physical and interpretational problems.

## 1 Introduction

At the classical level the accepted mathematical description of the four interactions at the basis of our understanding of nature (gravitational, electromagnetic, weak and strong; without or with the not yet experimentally verified supersymmetry between half-integer and integer spin fields, i.e. between fermions and bosons) , is based on action principles which, due to manifest Lorentz invariance, to local gauge invariance (minimal coupling) and/or diffeomorphism invariances, make use of singular Lagrangians implying the Dirac-Bergmann theory of constraints[1, 2] for their Hamiltonian formulation. While behind the gauge freedom of gauge theories proper there are Lie groups acting on some internal space so that the measurable quantities must be gauge invariant, the gauge freedom of theories invariant under diffeomorphism groups of the underlying spacetime (general relativity, string theory and reparametrization invariant systems of relativistic particles) concerns the arbitrariness for the observer in the choice of the definition of “what is space and/or time” (and relative times in the case of particles), i.e. of the definitory properties either of spacetime itself or of the measuring apparatuses. This is the classical mathematical background on which our understanding of the quantum field theory of electromagnetic, weak and strong interactions in the modern BRS formulation is based. The same is true for our attempts to build quantum gravity notwithstanding our actual incapacity to reconcile the influence of gravitational physics on the existence and formulation of spacetime concepts with the basic ideas of quantum theory, which requires a given absolute background spacetime.

Current research on electromagnetic, weak and strong interactions in special relativity, namely in Minkowski spacetime, has partially bypassed the problem by the covariant approach based on the BRS symmetry which, at least at the level of the algebra of infinitesimal gauge transformations, allows a regularization and renormalization of the relevant theories inside the framework of local quantum field theory (see for instance Ref.[3]). However, problems like the understanding of finite gauge transformations and of the associated moduli spaces, the Gribov ambiguity dependence on the choice of the function space for the fields and the gauge transformations, the confinement of quarks, the definition of relativistic bound states and how to put them among the asymptotic states, the nonlocality of charged states in quantum electrodynamics, not to speak of the foundational and practical problems posed by gravity, suggest that we should revisit the foundations of our

---

\*Email: lusanna@fi.infn.it

theories. It is not yet known whether we can understand which are the physical degrees of freedom hidden behind manifest gauge and/or general covariance and whether we can firstly meaningfully reformulate classical physics in terms of them and secondly to quantize the resulting theories. This will require to abandon local field theory at the nonperturbative level and to understand how to regularize and renormalize the Coulomb gauge of electrodynamics to start with. Moreover, the special relativistic theories will have to be reformulated in such a way to allow a natural transition to the coupling to gravity. Even if usually gravitational contributions are ignored because they are too weak with respect to the other interactions, the existing solution to the ultraviolet divergences of quantum field theory is distributional, so that, at least at the mathematical level, it is not justified to ignore gravity with all its nonlinearities. In turn general relativity must be formulated in a way allowing its deparametrization to recover physics in Minkowski spacetime when the Newton constant is put equal to zero. One also needs a formulation in which some notion of elementary particle exists so to recover Wigner's definition based on the irreducible representations of the Poincaré group in Minkowski spacetime with the further enrichment of the known good quantum numbers for their classification. Moreover, one needs some way out from the "problem of time" [4, 5, 6], since neither any consistent way to quantize time (is it a necessity?), and generically any timelike variable, nor a control on the associated problem of the relative times of a system of relativistic particles are known. Finally, one has to find a solution to the more basic problem of how to identify physically spacetime points in Einstein's formulation of general relativity, where general covariance deprives the mathematical points of the underlying 4-manifold of any physical reality [7, 8], while, on the experimental side (space physics, gravitational waves detectors), we are employing a theory of measurements of proper times and spacelike lengths which presupposes the individuation of points. This problem will appear also in the nowadays most popular program of unification of all the interactions in a supersymmetric way, i.e. in superstring theory and in its searched M-theory extension (see for instance Ref.[9]; string theory will not be touched in this review), when someone will be able to reformulate it in a background independent way.

These motivations induced me to revisit the classical Hamiltonian formulation of theories described by singular Lagrangians trying to choose the mathematical frameworks which at each step looked more natural to clarify the physical interpretational problems by means of the use of suitable adapted coordinates. In particular, after many years of dominance of the point of view privileging manifest Lorentz, gauge and general covariance at the price of losing control on the physical degrees of freedom and on their deterministic evolution (felt as a not necessary luxury only source of difficulties and complications), I went back to the old concept of Dirac observables, namely of those gauge invariant deterministic variables which describe a canonical basis of measurable quantities for the electromagnetic, weak and strong interactions in Minkowski spacetime. Instead, in general relativity, due to the problem of the individuation of the points of spacetime, measurable quantities have a more complex identification, which coincides with Dirac's observables (in any case indispensable for the treatment of the Cauchy problem) only in a completely fixed gauge (total breaking of general covariance).

In the next Sections I will review the various achievements of the program at the present stage of development (see Refs.[10] for previous reviews). Since there is too vast a bibliography to be covered in this review, I made the choice to concentrate it on my point of view omitting to quote many aspects of the theory and the work of many researchers.

## 1 Singular Lagrangians, Presymplectic Geometry, the Shanmugadhasan Canonical Transformations and Generalized Coulomb Gauges in Minkowski Spacetime.

A) If a finite-dimensional system with configuration space  $Q$  [ $q^i$ ,  $i=1,\dots,N$ , are local coordinates in a global (assumed to exist for the sake of simplicity) chart of the atlas of  $Q$ ;  $(t, q^i(t))$  is a point in  $R \times Q$ , where  $R$  is the time axis;  $\dot{q}^i(t) = dq^i(t)/dt$ ] is described by a singular Lagrangian  $L$  [so that the Hessian matrix is degenerate:  $\det(\partial^2 L / \partial \dot{q}^i \partial \dot{q}^j) = 0$ ], its Euler-Lagrange equations are in

general a mixture of equations i) depending only on the  $q^i$  (holonomic constraints); ii) depending only on  $q^i$  and  $\dot{q}^i$  (Lagrangian, in general nonholonomic, constraints and/or intrinsic first order equations of motion violating the so called second order differential equation (SODE) conditions); iii) depending on  $q^i$ ,  $\dot{q}^i$ ,  $\ddot{q}^i$  (genuine second order equations of motion, which however cannot be put in normal form, i.e. solved in the  $\ddot{q}^i$ ). More equations of the types i) and ii) can be deduced from the Euler-Lagrange equations and their time derivatives. The study of this type of degenerate equations can be traced back to Levi-Civita[11]. The solutions of the Euler-Lagrange equations depend on arbitrary functions of time, namely they are not deterministic.

The canonical momenta  $p_i = \partial L / \partial \dot{q}^i$  are not independent: there are relations among them  $\phi_\alpha(q, p) \approx 0$  called primary Hamiltonian constraints, which define a submanifold  $\gamma$  of the cotangent space  $T^*Q$  [the model is defined only on this submanifold; one uses the Poisson brackets of  $T^*Q$  in a neighbourhood of  $\gamma$  and Dirac's weak equality  $\approx$  means that the equality sign cannot be used inside Poisson brackets]. The canonical Hamiltonian  $H_c(q, p)$  has to be replaced by the Dirac Hamiltonian  $H_D = H_c + \sum_\alpha \lambda_\alpha(t) \phi_\alpha$ , which knows the restriction to the submanifold  $\gamma$  due to the arbitrary Dirac multipliers  $\lambda_\alpha(t)$ . The time constancy of the primary constraints,  $\partial_t \phi_\alpha = \{\phi_\alpha, H_D\} \approx 0$ , either produces secondary Hamiltonian constraints or determines some of the Dirac multipliers. This procedure is repeated for the secondary constraints (this is the Dirac-Bergmann algorithm) and so on. At the end there is a final set of constraints  $\chi_a \approx 0$  defining the final submanifold  $\bar{\gamma}$  of  $T^*Q$  on which the dynamics is consistently restricted, and a final Dirac Hamiltonian with a reduced set of arbitrary Dirac multipliers describing the remaining indetermination of the time evolution. The constraints are divided into two subgroups: i) the first class ones  $\chi_m^{(1)} \approx 0$ , having weakly zero Poisson bracket with all constraints and being the generators of the gauge transformations of the theory (the associated vector fields  $\{., \chi_m^{(1)}\}$  are tangent to  $\bar{\gamma}$ ); ii) the second class ones  $\chi_n^{(2)} \approx 0$  (their number is even) with  $\det(\{\chi_{n_1}^{(2)}, \chi_{n_2}^{(2)}\}) \neq 0$ , corresponding to pairs of inessential eliminable variables (the associated vector fields are normal to  $\bar{\gamma}$ ). The solutions of the Hamilton-Dirac equations with the final Dirac Hamiltonian depend on as many arbitrary functions of time as the left Dirac multipliers. The restriction of the symplectic 2-form of  $T^*Q$  to  $\bar{\gamma}$  is a closed degenerate 2-form, which in case of only first class constraints generates a so called presymplectic geometry:  $\bar{\gamma}$  is said to be a presymplectic manifold coisotropically embedded in  $T^*Q$  [see Ref.[12, 13] for what is known on presymplectic structures (they are dual to Poisson structures, but much less studied not being connected with integrable systems) and on the more general ones when second class constraints are present]. When many mathematical conditions are satisfied, the vector fields associated with the first class constraints (they are in the kernel of the degenerate 2-form on  $\bar{\gamma}$ ) generate a foliation of the submanifold  $\bar{\gamma}$ : each leaf (Hamiltonian gauge orbit) contains all the configurations which are gauge equivalent and which have to be considered as the same physical configuration[1] (equivalence class of gauge equivalent configurations); the canonical Hamiltonian  $H_c$  (if it is not  $H_c \approx 0$ ) generates an evolution which maps one leaf into the others. Therefore, the physical reduced phase space is obtained: i) by eliminating as many pairs of conjugate variables as second class constraints by means of the so called associated Dirac brackets; ii) by going to the quotient with respect to the foliation (a representative of the reduced phase space can be build by adding as many gauge-fixing constraints as first class ones, so to obtain a set of second class constraints). In general this procedure breaks the original Lorentz invariance.

Let us remark that only the primary first class constraints are associated with arbitrary Dirac multipliers. The secondary, tertiary... first class constraints are, in general, present in the canonical Hamiltonian  $H_c$  multiplied by well defined functions of  $q^i$ ,  $\dot{q}^i$ , which turn out to be arbitrary because they are not determined by the Hamilton-Dirac equations (they are gauge variables). This contradicts the Dirac conjecture[1] that the secondary first class constraints can be added to the Dirac Hamiltonian with extra multipliers (the resulting extended Dirac Hamiltonian would not allow the reconstruction of the original singular Lagrangian by inverse Legendre transformation; since the difference in the dynamics is only off-shell, this explains why the extended Hamiltonian is used in the BFV approach[14]). The natural way to add gauge-fixing constraints when there are secondary first class constraints[15], is to start giving the gauge fixings to the secondary constraints. The requirement of time constancy of these gauge fixings will generate the gauge fixings for the

primary first class constraints and the time constancy of these new gauge fixings will determine the Dirac multipliers eliminating every residual gauge freedom.

The Dirac observables are the gauge invariant functions on the reduced phase space, on which there is a deterministic evolution generated by the projection of the canonical Hamiltonian. Therefore, the main problem is to find a (possibly global) Darboux coordinate chart of the reduced phase space, namely a canonical basis of Dirac observables (or at least a Poisson algebra of them, according to Ref.[16]).

One would expect that when this is not possible, the relativistic system is intrinsically ill defined already at the classical level: at the quantum level this should manifest itself with the presence of not curable anomalies (which can be present also for a classically well defined system). Since the mathematical theory of the anomalies relies on cohomological properties of the manifolds (like  $Q$  and  $\bar{\gamma}$ ) relevant to the description of the system, which have to be defined already at the classical level, one expects that a classical background of these properties in the form of obstructions to the determination of the observables should be present in the theory of classical gauge canonical transformations.

When there is reparametrization invariance of the original action  $S = \int dt L$ , the canonical Hamiltonian vanishes and the reduced phase space is said frozen (like it happens in Hamilton-Jacobi theory). When the canonical Hamiltonian vanishes, both kinematics and dynamics are contained in the first class constraints describing the system: these can be interpreted as generalized Hamilton-Jacobi equations[17], so that the Dirac observables turn out to be the Jacobi data. When there is a kinematical symmetry group, like the Galileo or Poincaré groups, an evolution may be reintroduced by using the energy generator as Hamiltonian.

In a series of papers[18, 19, 20, 21, 22, 23] I made a reformulation of the general theory of singular Lagrangians and Hamiltonian constraints based on an extension of the second Noether theorem[24] to include also second class constraints. By means of the resulting Noether identities the Dirac-Bergmann algorithm was reproduced at the Lagrangian level. All the obscure and/or ambiguous points of the theory were clarified. The understanding[19] of the pathological examples known in the literature led to the discovery of third- and fourth-class constraints [with their associated singularities of the Jacobi equations (linearization of the Euler-Lagrange equations) and their connection with the reject of the Dirac conjecture about adding the secondary first class constraints to the Dirac Hamiltonian with extra Dirac multipliers] and of the phenomena of proliferation of constraints, ramification and joining of chains of constraints. Also the classification of all possible patterns of second class constraints was given[23]. All these phenomena have their counterpart in the study of the Euler-Lagrange equations for a singular Lagrangian in the second-order formalism. In Ref.[22] there is also the status of the art for the much more difficult and still incomplete first-order formulation of the theory on the tangent space  $TQ$  or on the first jet bundle  $J^1(Q) \approx TQ \times R$ , while in Ref.[21] there is the connection with BRS theory.

B) Now I will delineate the main steps for the determination of the Dirac observables for the case in which only primary first class constraints  $\phi_\alpha \approx 0$  are present at the Hamiltonian level.

The Euler-Lagrange equations associated with a singular Lagrangian do not determine the gauge part of the extremals. However it cannot be totally arbitrary, but must be compatible with the algebraic properties of the Noether gauge transformations induced by the first class constraints under which the action is either invariant or quasi-invariant as implied by the second Noether theorem. In the Hamiltonian formulation these properties are contained in the structure constants, or functions, of the Poisson brackets of the first-class constraints among themselves  $[\{\phi_\alpha, \phi_\beta\} = C_{\alpha\beta\gamma}\phi_\gamma, \{\phi_\alpha, H_c\} = C_{\alpha\beta}\phi_\beta]$  and the gauge arbitrariness of the trajectories is described by the Dirac multipliers appearing in the Dirac Hamiltonian. In both formulations one has to add extra equations, the either Lagrangian or Hamiltonian multitemporal equations[20], to have a consistent determination of the gauge part of the trajectory (see the generalized Lie equations of Ref.[25]). These equations are obtained by rewriting the variables  $q^i(t)$ ,  $p_i(t)$  in the form  $q^i(t, \tau_\alpha)$ ,  $p_i(t, \tau_\alpha)$ , and by assuming that the original  $t$ -evolution generated by the Dirac Hamiltonian  $H_D = H_c + \sum_\alpha \lambda_\alpha(t)\phi_\alpha$  is replaced by: i) a deterministic  $t$ -evolution generated by  $H_c$ ; ii) a  $\tau_\alpha$ -evolution (reabsorbing the arbitrary Dirac multipliers  $\lambda_\alpha(t)$ ), for each  $\alpha$ , generated in a suitable way by the first class constraints  $\phi_\alpha$ . The  $\tau_\alpha$ -dependence of  $q^i$ ,  $p_i$  determined by these multitemporal



(or better multiparametric) equations, which are integrable due to the first-class property of the constraints, describes their dependence on the gauge orbit containing the given Cauchy data for the Hamilton-Dirac equations. From the point of view of the study of the multitemporal equations, the secondary first class constraints are treated like the primary ones, namely as if there would be associated extra Dirac multipliers, and one should use as canonical Hamiltonian  $H_c$  restricted to zero value of the secondary constraints.

When the Poisson brackets of the Hamiltonian first class constraints imply a canonical realization of a Lie algebra, the extra Hamiltonian multitemporal equations have the first class constraints as Hamiltonians (so that the Dirac Hamiltonian is reduced to the canonical Hamiltonian) and the time parameters (replacing the Dirac multipliers) are the coordinates of a group manifold for a Lie group whose algebra is the given Lie algebra: they enter in the multitemporal equations via a set of left invariant vector fields  $Y_\alpha$  on the group manifold [ $Y_\alpha A(q, p) = \{A(q, p), \phi_\alpha\}$ ]. In the ideal case in which the gauge foliation of  $\bar{\gamma}$  is nice, all the leaves (or gauge orbits) are diffeomorphic and in the simplest case all of them are diffeomorphic to the group manifold of a Lie group. In this ideal case to rebuild a gauge orbit from one of its points (and therefore to determine the gauge part of the trajectories passing through that point) one needs the Lie equations associated with the given Lie group: the Hamiltonian multitemporal equations are generalized Lie equations describing all the gauge orbits simultaneously. In a generic case this description holds only locally for a set of diffeomorphic orbits, also in the case of systems invariant under diffeomorphisms.

Once one has solved the multitemporal equations, the next step is the determination of a Shanmugadhasan canonical transformation[26]. In the finite dimensional case general theorems[27] connected with the Lie theory of function groups[28] ensure the existence of local canonical transformations from the original canonical variables  $q^i, p_i$ , in terms of which the first class constraints (assumed globally defined) have the form  $\phi_\alpha(q, p) \approx 0$ , to canonical bases  $P_\alpha, Q_\alpha, P_A, Q_A$ , such that the equations  $P_\alpha \approx 0$  locally define the same original constraint manifold (the  $P_\alpha$  are an Abelianization of the first class constraints); the  $Q_\alpha$  are the adapted Abelian gauge variables describing the gauge orbits (they are a realization of the times  $\tau_\alpha$  of the multitemporal equations in terms of variables  $q^i, p_i$ ); the  $Q_A, P_A$  are an adapted canonical basis of Dirac observables. These canonical transformations are the basis of the Hamiltonian definition of the Faddeev-Popov measure of the path integral[29] and give a trivialization of the BRS construction of observables (the BRS method works when the first class constraints may be Abelianized[30]). Therefore the problem of the search of the Dirac observables becomes the problem of finding Shanmugadhasan canonical transformations. The strategy is to find abelianizations  $P_\alpha$  of the original constraints, to solve the multitemporal equations for  $q^i, p_i$  associated with the  $P_\alpha$ , to determine the multitudes  $Q_\alpha = \tau_\alpha$  and to identify the Dirac observables  $P_A, Q_A$  from the remaining original variables, i.e. from those their combinations independent from  $P_\alpha$  and  $Q_\alpha$ . Second class constraints, when present, are also taken into account by the Shanmugadhasan canonical transformation[26].

Putting equal to zero the Abelianized gauge variables one defines a local gauge of the model. If a system with constraints admits one (or more) global Shanmugadhasan canonical transformations, one obtains one (or more) privileged global gauges in which the physical Dirac observables are globally defined and globally separated from the gauge degrees of freedom [for systems with a compact configuration space  $Q$  this is impossible]. These privileged gauges (when they exist) can be called generalized Coulomb gauges. When the system under investigation has some global symmetry group, the associated theory of the momentum map[31] is a source of globality.

C) Now all the physical systems defined in the flat Minkowski spacetime, have the global Poincare' symmetry. This suggests to study the structure of the constraint manifold  $\bar{\gamma}$  from the point of view of the orbits of the Poincare' group. If  $p^\mu$  is the total momentum of the system, the constraint manifold has to be divided in four strata (some of them may be absent for certain systems) according to whether  $p^2 > 0$ ,  $p^2 = 0$ ,  $p^2 < 0$  or  $p^\mu = 0$ . Due to the different little groups of the various Poincare' orbits, the gauge orbits of different sectors will not be diffeomorphic. Therefore the manifold  $\bar{\gamma}$  is a stratified manifold and the gauge foliations of relativistic systems are nearly never nice, but rather one has to do with singular foliations.

For an acceptable relativistic system the stratum  $p^2 < 0$  has to be absent to avoid tachyons. To study the strata  $p^2 = 0$  and  $p^\mu = 0$  one has to add these relations as extra constraints. For all the

strata the next step (see however the next Section) is to do a canonical transformation from the original variables to a new set consisting of center-of-mass variables  $x^\mu$ ,  $p^\mu$  and of variables relative to the center of mass. Let us now consider the stratum  $p^2 > 0$ . By using the standard Wigner boost  $L^\mu_\nu(p, \hat{p})$  ( $p^\mu = L^\mu_\nu(p, \hat{p})\hat{p}^\nu$ ,  $\hat{p}^\mu = \eta\sqrt{p^2}(1; \vec{0})$ ,  $\eta = \text{sign } p^0$ ), one boosts the relative variables at rest. The new variables are still canonical and the base is completed by  $p^\mu$  and by a new center-of-mass coordinate  $\tilde{x}^\mu$ , differing from  $x^\mu$  for spin terms. The variable  $\tilde{x}^\mu$  has complicated covariance properties; instead the new relative variables are either Poincaré' scalars or Wigner spin-1 vectors, transforming under the group  $O(3)(p)$  of the Wigner rotations induced by the Lorentz transformations. A final canonical transformation[32], leaving fixed the relative variables, sends the center-of-mass coordinates  $\tilde{x}^\mu$ ,  $p^\mu$  in the new set  $p \cdot \tilde{x}/\eta\sqrt{p^2} = p \cdot x/\eta\sqrt{p^2}$  (the time in the rest frame),  $\eta\sqrt{p^2}$  (the total mass),  $\vec{k} = \vec{p}/\eta\sqrt{p^2}$  (the spatial components of the 4-velocity  $k^\mu = p^\mu/\eta\sqrt{p^2}$ ,  $k^2 = 1$ ),  $\vec{z} = \eta\sqrt{p^2}(\vec{\tilde{x}} - \tilde{x}^0 \vec{p}/p^0)$ .  $\vec{z}$  is a noncovariant center-of-mass canonical 3-coordinate multiplied by the total mass: it is the classical analog of the Newton-Wigner position operator (like it,  $\vec{z}$  is covariant only under the little group  $O(3)(p)$  of the timelike Poincaré orbits). Analogous considerations could be done for the other sectors. In Ref.[33] there is the definition of other canonical bases, the spin bases, adapted to the spin Casimir of the Poincaré group.

The nature of the relative variables depends on the system. The first class constraints, once rewritten in terms of the new variables, can be manipulated to find suitable global and Lorentz scalar Abelianizations. Usually there is a combination of the constraints which determines  $\eta\sqrt{p^2}$ , i.e. the mass spectrum, so that the time in the rest frame  $p \cdot x/\eta\sqrt{p^2}$  is the conjugated Lorentz scalar gauge variable. The other constraints eliminate some of the relative variables (in particular the relative energies for systems of interacting relativistic particles and the string): their conjugated coordinates (the relative times) are the other gauge variables: they are identified with a possible set of time parameters by the multitemporal equations. The Dirac observables (apart from the center-of-mass ones  $\vec{k}$  and  $\vec{z}$ ) have to be extracted from the remaining relative variables and the construction shows that they will be either Poincaré' scalars or Wigner covariant objects. In this way in each stratum preferred global Shanmugadhasan canonical transformations are identified, when no other kind of obstruction to globality is present inside the various strata.

D) In gauge field theories the situation is more complicated, because the theorems ensuring the existence of the Shanmugadhasan canonical transformation have not been extended to the infinite-dimensional case. One of the reasons is that some of the constraints can now be interpreted as elliptic equations and they can have zero modes. Let us consider the stratum  $p^2 > 0$  of free Yang-Mills theory as a prototype and its first class constraints, given by the Gauss laws and by the vanishing of the time components of the canonical momenta. The problem of the zero modes will appear as a singularity structure of the gauge foliation of the allowed strata, in particular of the stratum  $p^2 > 0$ . This phenomenon was discovered in Ref.[34] by studying the space of solutions of Yang-Mills and Einstein equations, which can be mapped onto the constraint manifold of these theories in their Hamiltonian description. It turns out that the space of solutions has a "cone over cone" structure of singularities: if we have a line of solutions with a certain number of symmetries, in each point of this line there is a cone of solutions with one less symmetry. In the Yang-Mills case the "gauge symmetries" of a gauge potential are connected with the generators of its stability group, i.e. with the subgroup of those special gauge transformations which leave invariant that gauge potential (this is the Gribov ambiguity for gauge potentials; there is also a more general Gribov ambiguity for field strengths, the "gauge copies" problem). Since the Gauss laws are the generators of the gauge transformations (and depend on the chosen gauge potential through the covariant derivative), this means that for a gauge potential with non trivial stability group those combinations of the Gauss laws corresponding to the generators of the stability group cannot be any more first class constraints, since they do not generate effective gauge transformations but special symmetry transformations. This problematics has still to be clarified, but it seems that in this case these components of the Gauss laws become third class constraints, which are not generators of true gauge transformations. This new kind of constraints was introduced in Refs.[19, 22] in the finite dimensional case as a result of the study of some examples, in which the Jacobi equations (the linearization of the Euler-Lagrange equations) are singular, i.e. some of their solutions are

not infinitesimal deviations between two neighbouring extremals of the Euler-Lagrange equations. This interpretation seems to be confirmed by the fact that the singularity structure discovered in Ref.[34] follows from the existence of singularities of the linearized Yang-Mills and Einstein equations. These problems are part of the Gribov ambiguity, which, as a consequence, induces an extremely complicated stratification and also singularities in each Poincaré stratum of  $\tilde{\gamma}$ .

Other possible sources of singularities of the gauge foliation of Yang-Mills theory in the stratum  $p^2 > 0$  may be: i) different classes of gauge potentials identified by different values of the field invariants; ii) the orbit structure of the rest frame (or Thomas) spin  $\vec{S}$ , identified by the Pauli-Lubanski Casimir  $W^2 = -p^2 \vec{S}^2$  of the Poincaré group.

The final outcome of this structure of singularities is that the reduced phase-space, i.e. the space of the gauge orbits, is in general a stratified manifold with singularities[16]. In the stratum  $p^2 > 0$  of the Yang-Mills theory these singularities survive the Wick rotation to the Euclidean formulation and it is not clear how the ordinary path integral approach and the associated BRS method can take them into account. The search of a global canonical basis of Dirac observables for each stratum of the space of the gauge orbits can give a definition of the measure of the phase space path integral, but at the price of a non polynomial Hamiltonian. Therefore, if it is not possible to eliminate the Gribov ambiguity (assuming that it is only a mathematical obstruction without any hidden physics), the existence of global Dirac observables for Yang-Mills theory is very problematic.

E) Firstly, inspired by Ref.[35] where a canonical basis of Dirac observables was found for the electromagnetic field interacting with a fermion field (whose Dirac observable is a fermion field dressed with a Coulomb cloud), the canonical reduction to noncovariant generalized Coulomb gauges, with the determination of the physical Hamiltonian as a function of a canonical basis of Dirac's observables, has been achieved for the following isolated systems (for them one asks that the 10 conserved generators of the Poincaré algebra are finite so to be able to use group theory; theories with external fields can only be recovered as limits in some parameter of a subsystem of the isolated system):

1) Relativistic particle mechanics. Its importance stems from the fact that quantum field theory has no particle interpretation: this is forced on it by means of the asymptotic states of the reduction formalism which correspond to the quantization of independent one-body systems described by relativistic mechanics [or relativistic pseudoclassical mechanics [36], when one adds Grassmann variables to describe the intrinsic spin]. Besides the scalar particle ( $p^2 - m^2 \approx 0$  or  $p^2 \approx 0$ ), one has control on: i) the pseudoclassical electron[37] ( $p_\mu \xi^\mu - m \xi_5 \approx 0$  or  $p_\mu \xi^\mu \approx 0$ , where  $\xi^\mu, \xi_5$  are Grassmann variables;  $p^2 - m^2 \approx 0$  or  $p^2 \approx 0$  are implied; after quantization the Dirac equation is reproduced); ii) the pseudoclassical neutrino[38] ( $p_\mu \xi^\mu + \frac{i}{3} \epsilon^{\mu\nu\rho\sigma} p_\mu \xi_\nu \xi_\rho \xi_\sigma \approx 0$ ,  $p^2 \approx 0$ , giving the Weyl particle wave equation  $p_\mu \gamma^\mu (1 - \gamma_5) \psi(x) = 0$  after quantization); iii) the pseudoclassical photon[39] ( $p^2 \approx 0$ ,  $p_\mu \theta^\mu \approx 0$ ,  $p_\mu \theta^{*\mu} \approx 0$ ,  $\theta_\mu^* \theta^\mu \approx 0$ , where  $\theta^\mu, \theta^{*\mu}$  are a pair of complex Grassmann four-vectors to describe helicity  $\pm 1$ ; after quantization one obtains the photon wave equations  $\square A^\mu(x) = 0$ ,  $\partial_\mu A^\mu(x) = 0$ ; the Berezin-Marinov Grassmann distribution function allows to recover the classical polarization matrix of classical light and, in quantization, the quantum polarization matrix with the Stokes parameters); iv) the vector particle or pseudoclassical massive photon[40] [ $p^2 - \mu^2 + (1 - \lambda) p_\mu \theta^{*\mu} p_\nu \theta^\nu \approx 0$ ,  $\theta_\mu^* \theta^\mu \approx 0$ , which, after quantization, reproduce the Proca-like wave equation  $(\square + \mu^2) A^\mu(x) - (1 - \lambda) \partial^\mu \partial_\nu A^\nu(x) = 0$ ].

The most important two-body system is the Droz-Vincent-Todorov-Komar model [41] with an arbitrary action-at-a-distance interaction instantaneous in the rest frame as shown by its energy-momentum tensor[42] [ $p_i^2 - m_i^2 + V(r_\perp^2) \approx 0$ ,  $i=1,2$ ,  $r_\perp^\mu = (\eta^{\mu\nu} - p^\mu p^\nu / p^2) r_\nu$ ,  $r^\mu = x_1^\mu - x_2^\mu$ ,  $p_\mu = p_{1\mu} + p_{2\mu}$ ]. This model has been completely understood both at the classical and quantum level [32]. Its study led to the identification of a class of canonical transformations (utilizing the standard Wigner boost for timelike Poincaré orbits) which allowed to understand how to define suitable center-of-mass and relative variables (in particular a suitable relative energy is determined by a combination of the two first class constraints, so that the relative time variable is a gauge variable), how to find a quasi-Shanmugadhasan canonical transformation adapted to the constraint determining the relative energy, how to separate the four, topologically disjointed, branches of the mass spectrum (it is determined by the other independent combination of the constraints; therefore,

there is a distinct Shanmugadhasan canonical transformation for each branch). At the quantum level it was possible to find four physical scalar products, compatible with both the resulting coupled wave equations (i.e. independent from the relative and the absolute rest-frame times): they have been found as generalization of the two existing scalar products of the Klein-Gordon equation: all of them are non-local even in the limiting free case and differ among themselves for the sign of the norm of states on different mass-branches. This example shows that the physical scalar product knows the functional form of the constraints.

The connection with the Bethe-Salpeter equation of the quantized model has been studied in Ref.[43], where it is shown that the constraint wave function can be obtained from the Bethe-Salpeter one by multiplication for a delta function containing the relative energy to exclude its spurious solutions (non physical excitations in the relative energy). The extension of the model to two pseudoclassical electrons and to an electron and a scalar has been done in Ref.[44], and the first was used to get good fits to meson spectra.

The previous canonical transformations were then extended to  $N$  free particles described by  $N$  mass-shell first class constraints  $p_i^2 - m_i^2 \approx 0$  [45]:  $N-1$  suitable relative energies are determined by  $N-1$  combinations of the constraints (so that the conjugate  $N-1$  relative times are gauge variables), while the remaining combination determines the  $2^N$  branches of the mass spectrum. The  $N$  gauge freedoms associated with these  $N$  combinations of the first class constraints are the freedom of the observer: i) in the choice of the time parameter to be used for the overall evolution of the isolated system; ii) in the choice of the description of the relative motions with any given delay among the pairs of particles.

In Ref.[46] 2- and  $N$ -body Newton mechanics was reformulated in a multitemporal way in terms of  $N$  first class constraints obtained from the relativistic ones in the limit  $c \rightarrow \infty$ . After a comparison with predictive mechanics, it was shown that the “no-interaction-theorem” (namely that the multitemporal configurational and canonical position coordinates of a particle coincide only in absence of interactions) exists also at the nonrelativistic level, being a property of the multitemporal description of particles and not of the kinematical symmetry group.

2) Both the open and closed Nambu string, after an initial study with light-cone coordinates, have been treated[47] along the lines of the two-body model in the stratum  $p^2 > 0$ . Both Abelian Lorentz scalar constraints and gauge variables have been found and globally decoupled, and a redundant set of Dirac’s observables  $[\vec{z}, \vec{k}, \vec{a}_n]$  has been found. It remains an open problem whether one can extract a global canonical basis of Dirac’s observables from the Wigner spin 1 vectors  $\vec{a}_n$ , which satisfy sigma-model-like constraints; if this basis exists, it would define the Liouville integrability of the Nambu string and would clarify whether there is any way to quantize it in four dimensions.

3) Yang-Mills theory with Grassmann-valued fermion fields [48] in the case of a trivial principal bundle over a fixed- $x^0$   $R^3$  slice of Minkowski spacetime with suitable Hamiltonian-oriented boundary conditions; this excludes monopole solutions (to have them, even if they have been not yet found experimentally, one needs a nontrivial bundle and a variational principle formulated on the bundle[49], because the gauge potentials on Minkowski spacetime are not globally defined) and, since  $R^3$  is not compactified, one has only winding number and no instanton number. After a discussion of the Hamiltonian formulation of Yang-Mills theory, of its group of gauge transformations and of the Gribov ambiguity, the theory has been studied in suitable weighted Sobolev spaces where the Gribov ambiguity is absent [50] and the global color charges are well defined. The global Dirac observables are the transverse quantities  $\vec{A}_{a\perp}(\vec{x}, x^0)$ ,  $\vec{E}_{a\perp}(\vec{x}, x^0)$  and fermion fields dressed with Yang-Mills (gluonic) clouds. The nonlocal and nonpolynomial (due to the presence of classical Wilson lines along flat geodesics) physical Hamiltonian has been obtained: it is nonlocal but without any kind of singularities, it has the correct Abelian limit if the structure constants are turned off, and it contains the explicit realization of the abstract Mitter-Viallet metric.

4) The Abelian and non-Abelian  $SU(2)$  Higgs models with fermion fields[51], where the symplectic decoupling is a refinement of the concept of unitary gauge. There is an ambiguity in the solutions of the Gauss law constraints, which reflects the existence of disjoint sectors of solutions of the Euler-Lagrange equations of Higgs models. The physical Hamiltonian and Lagrangian of the Higgs phase have been found; the self-energy turns out to be local and contains a local four-fermion

interaction.

5) The standard  $SU(3) \times SU(2) \times U(1)$  model of elementary particles[52] with Grassmann-valued fermion fields. The final reduced Hamiltonian contains nonlocal self-energies for the electromagnetic and color interactions, but “local ones” for the weak interactions implying the nonperturbative emergence of 4-fermions interactions.

F) When a good description of the system in terms of Dirac observables exists, one is going to face the problem of quantizing only the true physical degrees of freedom, which generically are nonlinear and nonlocal functions or functionals of the original variables. When a quantization is possible, there is a high probability to get a quantum theory inequivalent to that obtained by first quantizing the original variables and then making the reduction to the physical degrees of freedom at the quantum level (see for instance the BRS method).

With regards to field theory, this method has the drawback that generically the physical Hamiltonian, and therefore also the Lagrangian, is non polynomial in the physical degrees of freedom. Power counting methods cannot be used when looking for regularizations and renormalizations of the theory, and the advantages of a global control of the dynamics of physical quantities and of the possibility to check whether a model is classically well defined are destroyed by our present inability to solve these problems. The question, which puzzled both Dirac and Yukawa, reappears, whether it is possible to define an intrinsic ultraviolet cutoff and a regularization scheme independent from the power counting.

## 2 The Separation of the Center of Mass in Special Relativity, the Rest-Frame Instant Form of Dynamics and Wigner-Covariant Generalized Coulomb Gauges.

The next problem is how to covariantize these results valid in Minkowski spacetime with Cartesian coordinates. Again the starting point was given by Dirac[1] with his reformulation of classical field theory on spacelike hypersurfaces foliating Minkowski spacetime  $M^4$  [the foliation is defined by an embedding  $R \times \Sigma \rightarrow M^4$ ,  $(\tau, \vec{\sigma}) \mapsto z^{(\mu)}(\tau, \vec{\sigma}) \in \Sigma_\tau$ , with  $\Sigma$  an abstract 3-surface diffeomorphic to  $R^3$ , with  $\Sigma_\tau$  its copy embedded in  $M^4$  labelled by the value  $\tau$  (the Minkowski flat indices are  $(\mu)$ ; the scalar “time” parameter  $\tau$  labels the leaves of the foliation,  $\vec{\sigma}$  are curvilinear coordinates on  $\Sigma_\tau$  and  $(\tau, \vec{\sigma})$  are  $\Sigma_\tau$ -adapted holonomic coordinates for  $M^4$ ); this is the classical basis of Tomonaga-Schwinger quantum field theory]. In this way one gets a parametrized field theory with a covariant 3+1 splitting of Minkowski spacetime and already in a form suited to the transition to general relativity in its ADM canonical formulation (see also Ref.[53], where a theoretical study of this problem is done in curved spacetimes). The price is that one has to add as new independent configuration variables the embedding coordinates  $z^{(\mu)}(\tau, \vec{\sigma})$  of the points of the spacelike hypersurface  $\Sigma_\tau$  [the only ones carrying Lorentz indices] and then to define the fields on  $\Sigma_\tau$  so that they know the hypersurface  $\Sigma_\tau$  of  $\tau$ -simultaneity [for a Klein-Gordon field  $\phi(x)$ , this new field is  $\tilde{\phi}(\tau, \vec{\sigma}) = \phi(z(\tau, \vec{\sigma}))$ : it contains the nonlocal information about the embedding]. Then one rewrites the Lagrangian of the given isolated system in the form required by the coupling to an external gravitational field, makes the previous 3+1 splitting of Minkowski spacetime and interpretes all the fields of the system as the new fields on  $\Sigma_\tau$  (they are Lorentz scalars, having only surface indices). Instead of considering the 4-metric as describing a gravitational field (and therefore as an independent field as it is done in metric gravity, where one adds the Hilbert action to the action for the matter fields), here one replaces the 4-metric with the induced metric  $g_{AB}[z] = z_A^{(\mu)} \eta_{(\mu)(\nu)} z_B^{(\nu)}$  on  $\Sigma_\tau$  [a functional of  $z^{(\mu)}$ ; here we use the notation  $\sigma^A = (\tau, \sigma^r)$ ;  $z_A^{(\mu)} = \partial z^{(\mu)} / \partial \sigma^A$  are flat tetrad fields on Minkowski spacetime with the  $z_r^{(\mu)}$ ’s tangent to  $\Sigma_\tau$ ] and considers the embedding coordinates  $z^{(\mu)}(\tau, \vec{\sigma})$  as independent fields [this is not possible in metric gravity, because in curved spacetimes  $z_A^\mu \neq \partial z^\mu / \partial \sigma^A$  are not tetrad fields so that holonomic coordinates  $z^\mu(\tau, \vec{\sigma})$  do not exist]. From this Lagrangian, besides a Lorentz-scalar form of the constraints of the given system, we get four extra primary first class constraints

$$\mathcal{H}_{(\mu)}(\tau, \vec{\sigma}) = \rho_{(\mu)}(\tau, \vec{\sigma}) - l_{(\mu)}(\tau, \vec{\sigma}) T_{sys}^{\tau\tau}(\tau, \vec{\sigma}) - z_{r(\mu)}(\tau, \vec{\sigma}) T_{sys}^{\tau r}(\tau, \vec{\sigma}) \approx 0$$

[here  $T_{sys}^{\tau\tau}(\tau, \vec{\sigma})$ ,  $T_{sys}^{\tau r}(\tau, \vec{\sigma})$ , are the components of the energy-momentum tensor in the holonomic coordinate system, corresponding to the energy- and momentum-density of the isolated system; one has  $\{\mathcal{H}_{(\mu)}(\tau, \vec{\sigma}), \mathcal{H}_{(\nu)}(\tau, \vec{\sigma})\} = 0$ ] implying the independence of the description from the choice of the 3+1 splitting, i.e. from the choice of the foliation with spacelike hypersurfaces. The evolution vector is given by  $z_r^{(\mu)} = N_{[z](flat)} l^{(\mu)} + N_{[z](flat)}^r z_r^{(\mu)}$ , where  $l^{(\mu)}(\tau, \vec{\sigma})$  is the normal to  $\Sigma_\tau$  in  $z^{(\mu)}(\tau, \vec{\sigma})$  and  $N_{[z](flat)}(\tau, \vec{\sigma})$ ,  $N_{[z](flat)}^r(\tau, \vec{\sigma})$  are the flat lapse and shift functions defined through the metric like in general relativity: however, now they are not independent variables but functionals of  $z^{(\mu)}(\tau, \vec{\sigma})$ .

The Dirac Hamiltonian contains the piece  $\int d^3\sigma \lambda^{(\mu)}(\tau, \vec{\sigma}) \mathcal{H}_{(\mu)}(\tau, \vec{\sigma})$  with  $\lambda^{(\mu)}(\tau, \vec{\sigma})$  Dirac multipliers. It is possible to rewrite the integrand in the form  $[\int d^3\sigma g^{rs}$  is the inverse of  $g_{rs}]$

$$\begin{aligned} \lambda_{(\mu)}(\tau, \vec{\sigma}) \mathcal{H}^{(\mu)}(\tau, \vec{\sigma}) &= [(\lambda_{(\mu)} l^{(\mu)})(l_{(\nu)} \mathcal{H}^{(\nu)}) - (\lambda_{(\mu)} z_r^{(\mu)}) ({}^3g^{rs} z_{s(\nu)} \mathcal{H}^{(\nu)})](\tau, \vec{\sigma}) \\ &\stackrel{def}{=} N_{(flat)}(\tau, \vec{\sigma}) (l_{(\mu)} \mathcal{H}^{(\mu)})(\tau, \vec{\sigma}) - N_{(flat)r}(\tau, \vec{\sigma}) ({}^3g^{rs} z_{s(\nu)} \mathcal{H}^{(\nu)})(\tau, \vec{\sigma}) \end{aligned}$$

with the (nonholonomic form of the) constraints  $(l_{(\mu)} \mathcal{H}^{(\mu)})(\tau, \vec{\sigma}) \approx 0$ ,  $({}^3g^{rs} z_{s(\mu)} \mathcal{H}^{(\mu)})(\tau, \vec{\sigma}) \approx 0$ , satisfying the universal Dirac algebra of the ADM constraints. In this way we have defined new flat lapse and shift functions

$$\begin{aligned} N_{(flat)}(\tau, \vec{\sigma}) &= \lambda_{(\mu)}(\tau, \vec{\sigma}) l^{(\mu)}(\tau, \vec{\sigma}), \\ N_{(flat)r}(\tau, \vec{\sigma}) &= \lambda_{(\mu)}(\tau, \vec{\sigma}) z_r^{(\mu)}(\tau, \vec{\sigma}). \end{aligned}$$

which have the same content of the arbitrary Dirac multipliers  $\lambda_{(\mu)}(\tau, \vec{\sigma})$ , namely they multiply primary first class constraints satisfying the Dirac algebra. In Minkowski spacetime they are quite distinct from the previous lapse and shift functions  $N_{[z](flat)}$ ,  $N_{[z](flat)r}$ , defined starting from the metric. Instead in general relativity the lapse and shift functions defined starting from the 4-metric are the coefficients (in the canonical part  $H_c$  of the Hamiltonian) of secondary first class constraints satisfying the Dirac algebra.

In special relativity, it is convenient to restrict ourselves to arbitrary spacelike hyperplanes  $z^{(\mu)}(\tau, \vec{\sigma}) = x_s^{(\mu)}(\tau) + b_r^{(\mu)}(\tau) \sigma^r$ . Since they are described by only 10 variables, after this restriction we remain only with 10 first class constraints determining the 10 variables conjugate to the hyperplane in terms of the variables of the system:

$$\mathcal{H}^{(\mu)}(\tau) = p_s^{(\mu)} - p_{(sys)}^{(\mu)} \approx 0, \quad \mathcal{H}^{(\mu)(\nu)}(\tau) = S_s^{(\mu)(\nu)} - S_{(sys)}^{(\mu)(\nu)} \approx 0.$$

After the restriction to spacelike hyperplanes the previous piece of the Dirac Hamiltonian is reduced to  $\tilde{\lambda}^{(\mu)}(\tau) \mathcal{H}_{(\mu)}(\tau) - \frac{1}{2} \tilde{\lambda}^{(\mu)(\nu)}(\tau) \mathcal{H}_{(\mu)(\nu)}(\tau)$ . Since at this stage we have  $z_r^{(\mu)}(\tau, \vec{\sigma}) \approx b_r^{(\mu)}(\tau)$ , so that  $z_r^{(\mu)}(\tau, \vec{\sigma}) \approx N_{[z](flat)}(\tau, \vec{\sigma}) l^{(\mu)}(\tau, \vec{\sigma}) + N_{[z](flat)}^r(\tau, \vec{\sigma}) b_r^{(\mu)}(\tau, \vec{\sigma}) \approx \dot{x}_s^{(\mu)}(\tau) + \dot{b}_r^{(\mu)}(\tau) \sigma^r = -\tilde{\lambda}^{(\mu)}(\tau) - \tilde{\lambda}^{(\mu)(\nu)}(\tau) b_{r(\nu)}(\tau) \sigma^r$ , it is only now that we get the coincidence of the two definitions of flat lapse and shift functions (this point was missed in the older treatments of parametrized Minkowski theories):

$$\begin{aligned} N_{[z](flat)}(\tau, \vec{\sigma}) &\approx N_{(flat)}(\tau, \vec{\sigma}) = -\tilde{\lambda}_{(\mu)}(\tau) l^{(\mu)} - l^{(\mu)} \tilde{\lambda}_{(\mu)(\nu)}(\tau) b_s^{(\nu)}(\tau) \sigma^s, \\ N_{[z](flat)r}(\tau, \vec{\sigma}) &\approx N_{(flat)r}(\tau, \vec{\sigma}) = -\tilde{\lambda}_{(\mu)}(\tau) b_r^{(\mu)}(\tau) - b_r^{(\mu)}(\tau) \tilde{\lambda}_{(\mu)(\nu)}(\tau) b_s^{(\nu)}(\tau) \sigma^s. \end{aligned}$$

The 20 variables for the phase space description of a hyperplane are:

- i)  $x_s^{(\mu)}(\tau)$ ,  $p_s^{(\mu)}$ , parametrizing the origin of the coordinates on the family of spacelike hyperplanes. The four constraints  $\mathcal{H}^{(\mu)}(\tau) \approx 0$  say that  $p_s^{(\mu)}$  is determined by the 4-momentum of the isolated system.
- ii)  $b_A^{(\mu)}(\tau)$  (with the  $b_r^{(\mu)}(\tau)$ 's being three orthogonal spacelike unit vectors generating the fixed



$\tau$ -independent timelike unit normal  $b_\tau^{(\mu)} = l^{(\mu)}$  to the hyperplanes) and  $S_s^{(\mu)(\nu)} = -S_s^{(\nu)(\mu)}$  with the orthonormality constraints  $b_A^{(\mu)} {}^4\eta_{(\mu)(\nu)} b_B^{(\nu)} = {}^4\eta_{AB}$  [enforced by assuming the Dirac brackets  $\{S_s^{(\mu)(\nu)}, b_A^{(\rho)}\} = {}^4\eta^{(\rho)(\nu)} b_A^{(\mu)} - {}^4\eta^{(\rho)(\mu)} b_A^{(\nu)}$ ,  $\{S_s^{(\mu)(\nu)}, S_s^{(\alpha)(\beta)}\} = C_{(\gamma)(\delta)}^{(\mu)(\nu)(\alpha)(\beta)} S_s^{(\gamma)(\delta)}$  with  $C_{(\gamma)(\delta)}^{(\mu)(\nu)(\alpha)(\beta)}$  the structure constants of the Lorentz algebra]. In these variables there are hidden six independent pairs of degrees of freedom. The six constraints  $\mathcal{H}^{(\mu)(\nu)}(\tau) \approx 0$  say that  $S_s^{(\mu)(\nu)}$  coincides the spin tensor of the isolated system. Then one has that  $p_s^{(\mu)}$ ,  $J_s^{(\mu)(\nu)} = x_s^{(\mu)} p_s^{(\nu)} - x_s^{(\nu)} p_s^{(\mu)} + S_s^{(\mu)(\nu)}$ , satisfy the algebra of the Poincaré group.

Let us remark that, for each configuration of an isolated system there is a privileged family of hyperplanes (the Wigner hyperplanes orthogonal to  $p_s^{(\mu)}$ , existing when  $p_s^2 > 0$ ) corresponding to the intrinsic rest-frame of the isolated system. If we choose these hyperplanes with suitable gauge fixings, we remain with only the four constraints  $\mathcal{H}^{(\mu)}(\tau) \approx 0$ , which can be rewritten as

$$\sqrt{p_s^2} \approx [\text{invariant mass of the isolated system under investigation}] = M_{sys};$$

$$\vec{p}_{sys} = [3 - \text{momentum of the isolated system inside the Wigner hyperplane}] \approx 0.$$

There is no more a restriction on  $p_s^{(\mu)}$ , because  $u_s^{(\mu)}(p_s) = p_s^{(\mu)}/p_s^2$  gives the orientation of the Wigner hyperplanes containing the isolated system with respect to an arbitrary given external observer.

In this special gauge we have  $b_A^{(\mu)} \equiv L^{(\mu)}_{A(p_s, \vec{p}_s)}$  (the standard Wigner boost for timelike Poincaré orbits),  $S_s^{(\mu)(\nu)} \equiv S_{system}^{(\mu)(\nu)}$ , and the only remaining canonical variables are the noncovariant Newton-Wigner-like canonical “external” center-of-mass coordinate  $\tilde{x}_s^{(\mu)}(\tau)$  (living on the Wigner hyperplanes) and  $p_s^{(\mu)}$ . Now 3 degrees of freedom of the isolated system [an “internal” center-of-mass 3-variable  $\vec{\sigma}_{sys}$  defined inside the Wigner hyperplane and conjugate to  $\vec{p}_{sys}$ ] become gauge variables [the natural gauge fixing is  $\vec{\sigma}_{sys} \approx 0$ , so that it coincides with the origin  $x_s^{(\mu)}(\tau) = z^{(\mu)}(\tau, \vec{\sigma} = 0)$  of the Wigner hyperplane], while the  $\tilde{x}^{(\mu)}$  is playing the role of a kinematical external center of mass for the isolated system and may be interpreted as a decoupled observer with his parametrized clock (point particle clock). All the fields living on the Wigner hyperplane are now either Lorentz scalar or with their 3-indices transforming under Wigner rotations (induced by Lorentz transformations in Minkowski spacetime) as any Wigner spin 1 index.

One obtains in this way a new kind of instant form of the dynamics (see Ref.[54]), the “Wigner-covariant 1-time rest-frame instant form”[55] with a universal breaking of Lorentz covariance. It is the special relativistic generalization of the nonrelativistic separation of the center of mass from the relative motion [ $H = \frac{\vec{P}^2}{2M} + H_{rel}$ ]. The role of the center of mass is taken by the Wigner hyperplane, identified by the point  $\tilde{x}^{(\mu)}(\tau)$  and by its normal  $p_s^{(\mu)}$ . The invariant mass  $M_{sys}$  of the system replaces the nonrelativistic Hamiltonian  $H_{rel}$  for the relative degrees of freedom, after the addition of the gauge-fixing  $T_s - \tau \approx 0$  [identifying the time parameter  $\tau$ , labelling the leaves of the foliation, with the Lorentz scalar time of the center of mass in the rest frame,  $T_s = p_s \cdot \tilde{x}_s / M_{sys}$ ;  $M_{sys}$  generates the evolution in this time].

The determination of  $\vec{\sigma}_{sys}$  may be done with the group theoretical methods of Ref.[56]: given a realization on the phase space of a given system of the ten Poincaré generators one can build three 3-position variables only in terms of them, which in our case of a system on the Wigner hyperplane with  $\vec{p}_{sys} \approx 0$  are: i) a canonical center of mass (the “internal” center of mass  $\vec{\sigma}_{sys}$ ); ii) a noncanonical Møller center of energy  $\vec{\sigma}_{sys}^{(E)}$ ; iii) a noncanonical Fokker-Pryce center of inertia  $\vec{\sigma}_{sys}^{(FP)}$ . Due to  $\vec{p}_{sys} \approx 0$ , we have  $\vec{\sigma}_{sys} \approx \vec{\sigma}_{sys}^{(E)} \approx \vec{\sigma}_{sys}^{(FP)}$ . By adding the gauge fixings  $\vec{\sigma}_{sys} \approx 0$  one can show that the origin  $x_s^{(\mu)}(\tau)$  becomes simultaneously the Dixon center of mass of an extended object and both the Pirani and Tulczyjew centroids (see Ref. [57] for the application of these methods to find the center of mass of a configuration of the Klein-Gordon field after the preliminary work of Ref.[58]). With similar methods one can construct three “external” collective positions (all located on the Wigner hyperplane): i) the “external” canonical noncovariant center of mass  $\tilde{x}_s^{(\mu)}$ ; ii) the “external” noncanonical and noncovariant Møller center of energy  $R_s^{(\mu)}$ ; iii) the “external” covariant noncanonical Fokker-Pryce center of inertia  $Y_s^{(\mu)}$  (when there are the gauge

fixings  $\vec{\sigma}_{sys} \approx 0$  it also coincides with the origin  $x_s^{(\mu)}$ . It turns out that the Wigner hyperplane is the natural setting for the study of the Dixon multipoles of extended relativistic systems[59] and for defining the canonical relative variables with respect to the center of mass. After having put control on the relativistic definitions of center of mass of an extended system, the lacking kinematics of relativistic rotations is now under investigation. The Wigner hyperplane with its natural Euclidean metric structure offers a natural solution to the problem of boost for lattice gauge theories and realizes explicitly the machian aspect of dynamics that only relative motions are relevant.

The isolated systems till now analyzed to get their rest-frame Wigner-covariant generalized Coulomb gauges [i.e. the subset of global Shanmugadhasan canonical bases, which, for each Poincaré stratum, are also adapted to the geometry of the corresponding Poincaré orbits with their little groups; these special bases can be named Poincaré-Shanmugadhasan bases for the given Poincaré stratum of the presymplectic constraint manifold (every stratum requires an independent canonical reduction); till now only the main stratum with  $p^2$  timelike and  $W^2 \neq 0$  has been investigated] are:

a) The system of  $N$  scalar particles with Grassmann electric charges plus the electromagnetic field [55]. The starting configuration variables are a 3-vector  $\vec{\eta}_i(\tau)$  for each particle [ $x_i^{(\mu)}(\tau) = z^{(\mu)}(\tau, \vec{\eta}_i(\tau))$ ] and the electromagnetic gauge potentials  $A_A(\tau, \vec{\sigma}) = \frac{\partial z^{(\mu)}(\tau, \vec{\sigma})}{\partial \sigma^A} A_{(\mu)}(z(\tau, \vec{\sigma}))$ , which know the embedding of  $\Sigma_\tau$  into  $M^4$ . One has to choose the sign of the energy of each particle, because there are not mass-shell constraints (like  $p_i^2 - m_i^2 \approx 0$ ) among the constraints of this formulation, due to the fact that one has only three degrees of freedom for particle, determining the intersection of a timelike trajectory and of the spacelike hypersurface  $\Sigma_\tau$ . For each choice of the sign of the energy of the  $N$  particles, one describes only one of the  $2^N$  branches of the mass spectrum of the manifestly covariant approach based on the coordinates  $x_i^{(\mu)}(\tau)$ ,  $p_i^{(\mu)}(\tau)$ ,  $i=1, \dots, N$ , and on the constraints  $p_i^2 - m_i^2 \approx 0$  (in the free case). In this way, one gets a description of relativistic particles with a given sign of the energy with consistent couplings to fields and valid independently from the quantum effect of pair production [in the manifestly covariant approach, containing all possible branches of the particle mass spectrum, the classical counterpart of pair production is the intersection of different branches deformed by the presence of interactions]. The final Dirac's observables are: i) the transverse radiation field variables  $\vec{A}_\perp, \vec{E}_\perp$ ; ii) the particle canonical variables  $\vec{\eta}_i(\tau), \vec{\pi}_i(\tau)$ , dressed with a Coulomb cloud. The physical Hamiltonian contains the mutual instantaneous Coulomb potentials extracted from field theory and there is a regularization of the Coulomb self-energies due to the Grassmann character of the electric charges  $Q_i$  [ $Q_i^2 = 0$ ]. In Ref.[60] there is the study of the Lienard-Wiechert potentials and of Abraham-Lorentz-Dirac equations in this rest-frame Coulomb gauge and also scalar electrodynamics is reformulated in it. Also the rest-frame 1-time relativistic statistical mechanics has been developed [55].

b) The system of  $N$  scalar particles with Grassmann-valued color charges plus the color  $SU(3)$  Yang-Mills field[61]: it gives the pseudoclassical description of the relativistic scalar-quark model, deduced from the classical QCD Lagrangian and with the color field present. The physical invariant mass of the system is given in terms of the Dirac observables. From the reduced Hamilton equations the second order equations of motion both for the reduced transverse color field and the particles are extracted. Then, one studies the  $N=2$  (meson) case. A special form of the requirement of having only color singlets, suited for a field-independent quark model, produces a "pseudoclassical asymptotic freedom" and a regularization of the quark self-energy. With these results one can covariantize the bosonic part of the standard model given in Ref.[52].

c) The system of  $N$  spinning particles of definite energy  $[(\frac{1}{2}, 0)$  or  $(0, \frac{1}{2})$  representation of  $SL(2, C)$ ] with Grassmann electric charges plus the electromagnetic field[62] and that of a Grassmann-valued Dirac field plus the electromagnetic field (the pseudoclassical basis of QED) [63]. In both cases there are geometrical complications connected with the spacetime description of the path of electric currents and not only of their spin structure, suggesting a reinterpretation of the supersymmetric scalar multiplet as a spin fibration with the Dirac field in the fiber and the Klein-Gordon field in the base; a new canonical decomposition of the Klein-Gordon field into center-of-mass and relative variables [58, 57] will be helpful to clarify these problems. After their solution and after



having obtained the description of Grassmann-valued chiral fields [this will require the transcription of the front form of the dynamics in the instant one for the Poincaré strata with  $P^2 = 0$ ] the rest-frame form of the full standard  $SU(3) \times SU(2) \times U(1)$  model can be achieved.

The rest-frame description of the relativistic perfect gas is now under investigation.

All these new pieces of information will allow, after quantization of this new consistent relativistic mechanics without the classical problems connected with pair production, to find the asymptotic states of the covariant Tomonaga-Schwinger formulation of quantum field theory on spacelike hypersurfaces (to be obtained by quantizing the fields on  $\Sigma_\tau$ ): these states are needed for the theory of quantum bound states [since Fock states do not constitute a Cauchy problem for the field equations, because an in (or out) particle can be in the absolute future of another one due to the tensor product nature of these asymptotic states, bound state equations like the Bethe-Salpeter one have spurious solutions which are excitations in relative energies, the variables conjugate to relative times]. Moreover, it will be possible to include bound states among the asymptotic states.

As said in Ref.[60, 61], the quantization of these rest-frame models has to overcome two problems. On the particle side, the complication is the quantization of the square roots associated with the relativistic kinetic energy terms: in the free case this has been done in Ref.[64] [see Refs.[65] for the complications induced by the Coulomb potential]. On the field side (all physical Hamiltonian are nonlocal and, with the exception of the Abelian case, nonpolynomial, but quadratic in the momenta), the obstacle is the absence (notwithstanding there is no no-go theorem) of a complete regularization and renormalization procedure of electrodynamics (to start with) in the Coulomb gauge: see Ref.[66] (and its bibliography) for the existing results for QED.

However, as shown in Refs.[55, 48], the rest-frame instant form of dynamics automatically gives a physical ultraviolet cutoff in the spirit of Dirac and Yukawa: it's the Møller radius[67]  $\rho = \sqrt{-W^2}/p^2 = |\vec{S}|/\sqrt{p^2}$  ( $W^2 = -p^2 \vec{S}^2$  is the Pauli-Lubanski Casimir when  $p^2 > 0$ ), namely the classical intrinsic radius of the worldtube, around the covariant noncanonical Fokker-Pryce center of inertia  $Y^{(\mu)}$ , inside which the noncovariance of the canonical center of mass  $\tilde{x}^\mu$  is concentrated. At the quantum level  $\rho$  becomes the Compton wavelength of the isolated system multiplied its spin eigenvalue  $\sqrt{s(s+1)}$ ,  $\rho \mapsto \hat{\rho} = \sqrt{s(s+1)}\hbar/M = \sqrt{s(s+1)}\lambda_M$  with  $M = \sqrt{p^2}$  the invariant mass and  $\lambda_M = \hbar/M$  its Compton wavelength. Therefore, the criticism to classical relativistic physics, based on quantum pair production, concerns the testing of distances where, due to the Lorentz signature of spacetime, one has intrinsic classical covariance problems: it is impossible to localize the canonical center of mass  $\tilde{x}^\mu$  adapted to the first class constraints of the system (also named Pryce center of mass and having the same covariance of the Newton-Wigner position operator) in a frame independent way.

Let us remember [55] that  $\rho$  is also a remnant in flat Minkowski spacetime of the energy conditions of general relativity: since the Møller noncanonical, noncovariant center of energy  $R^{(\mu)}$  has its noncovariance localized inside the same worldtube with radius  $\rho$  (it was discovered in this way) [67], it turns out that for an extended relativistic system with the material radius smaller of its intrinsic radius  $\rho$  one has: i) its peripheral rotation velocity can exceed the velocity of light; ii) its classical energy density cannot be positive definite everywhere in every frame.

Now, the real relevant point is that this ultraviolet cutoff determined by  $\rho$  exists also in Einstein's general relativity (which is not power counting renormalizable) in the case of asymptotically flat spacetimes, taking into account the Poincaré Casimirs of its asymptotic ADM Poincaré charges (when supertranslations are eliminated with suitable boundary conditions). The generalization of the worldtube of radius  $\rho$  to asymptotically flat general relativity with matter, could also be connected with the unproved cosmic censorship hypothesis.

Moreover, the extended Heisenberg relations of string theory[68], i.e.  $\Delta x = \frac{\hbar}{\Delta p} + \frac{\Delta p}{T_{cs}} = \frac{\hbar}{\Delta p} + \frac{\hbar \Delta p}{L_{cs}^2}$  implying the lower bound  $\Delta x > L_{cs} = \sqrt{\hbar/T_{cs}}$  due to the  $y + 1/y$  structure, have a counterpart in the quantization of the Møller radius[55]: if we ask that, also at the quantum level, one cannot test the inside of the worldtube, we must ask  $\Delta x > \hat{\rho}$  which is the lower bound implied by the modified uncertainty relation  $\Delta x = \frac{\hbar}{\Delta p} + \frac{\hbar \Delta p}{\hat{\rho}^2}$ . This could imply that the center-of-mass canonical noncovariant 3-coordinate  $\vec{z} = \sqrt{P^2}(\vec{x} - \frac{\vec{P}}{P^0} \tilde{x}^0)$  [55] cannot become a self-adjoint operator. See Hegerfeldt's theorems (quoted in Refs.[48, 55]) and his interpretation pointing at

the impossibility of a good localization of relativistic particles (experimentally one determines only a worldtube in spacetime emerging from the interaction region). Since the eigenfunctions of the canonical center-of-mass operator are playing the role of the wave function of the universe, one could also say that the center-of-mass variable has not to be quantized, because it lies on the classical macroscopic side of Copenhagen's interpretation and, moreover, because, in the spirit of Mach's principle that only relative motions can be observed, no one can observe it (it is only used to define a decoupled "point particle clock"). On the other hand, if one rejects the canonical noncovariant center of mass in favor of the covariant noncanonical Fokker-Pryce center of inertia  $Y^\mu$ ,  $\{Y^\mu, Y^\nu\} \neq 0$ , one could invoke the philosophy of quantum groups to quantize  $Y^\mu$  to get some kind of quantum plane for the center-of-mass description. Let us remark that the quantization of the square root Hamiltonian done in Ref.[64] is consistent with this problematic.

In conclusion, the best set of canonical coordinates adapted to the constraints and to the geometry of Poincaré orbits in Minkowski spacetime and naturally predisposed to the coupling to canonical tetrad gravity is emerging for the electromagnetic, weak and strong interactions with matter described either by fermion fields or by relativistic particles with a definite sign of the energy.

### 3 Tetrad Gravity, Physical Hamiltonian Degrees of Freedom of the Gravitational Field and the Deparametrization of General Relativity.

Tetrad gravity is the formulation of general relativity natural for the coupling to the fermion fields of the standard model. However, we need a formulation of it, which allows to solve its constraints for doing the canonical reduction and to solve the deparametrization problem of general relativity (how to recover the rest-frame instant form when the Newton constant is put equal to zero,  $G=0$ ). Since neither a complete reduction of gravity with an identification of the physical canonical degrees of freedom of the gravitational field nor a detailed study of its Hamiltonian group of gauge transformations (whose infinitesimal generators are the first class constraints) has ever been pushed till the end in an explicit way, a new formulation of tetrad gravity [69, 70, 71, 72] was developed.

To implement this program we shall restrict ourselves to the simplest class of spacetimes [time-oriented pseudo-Riemannian or Lorentzian 4-manifold  $(M^4, {}^4g)$  with signature  $\epsilon (+ - - -)$  ( $\epsilon = \pm 1$  according to either particle physics or general relativity convention) and with a choice of time orientation], assumed to be:

i) Globally hyperbolic 4-manifolds, i.e. topologically they are  $M^4 = R \times \Sigma$ , so to have a well posed Cauchy problem [with  $\Sigma$  the abstract model of Cauchy surface] at least till when no singularity develops in  $M^4$  [see the singularity theorems]. Therefore, these spacetimes admit regular foliations with orientable, complete, non-intersecting spacelike 3-manifolds  $\Sigma_\tau$  [ $\tau : M^4 \rightarrow R$ ,  $z^\mu \mapsto \tau(z^\mu)$ , is a global timelike future-oriented function labelling the leaves (surfaces of simultaneity)]. In this way, one obtains 3+1 splittings of  $M^4$  and the possibility of a Hamiltonian formulation.

ii) Asymptotically flat at spatial infinity, so to have the possibility to define asymptotic Poincaré charges [73, 74, 75]: they allow the definition of a Møller radius also in general relativity and are a bridge towards a future soldering with the theory of elementary particles in Minkowski spacetime defined as irreducible representation of its kinematical, globally implemented Poincaré group according to Wigner. This excludes Einstein-Wheeler closed universes without boundaries (no asymptotic Poincaré charges), which were introduced to eliminate boundary conditions at spatial infinity to make the theory as machian as possible.

iii) Admitting a spinor (or spin) structure[76] for the coupling to fermion fields. Since we consider noncompact space- and time-orientable spacetimes, spinors can be defined if and only if they are "parallelizable" [77], like in our case. This implies that the orthonormal frame principal  $SO(3)$ -bundle over  $\Sigma_\tau$  (whose connections are the spin connections determined by the cotriads) is trivial.

iv) The noncompact parallelizable simultaneity 3-manifolds (the Cauchy surfaces)  $\Sigma_\tau$  are assumed to be topologically trivial, geodesically complete and, finally, diffeomorphic to  $R^3$ . These 3-manifolds have the same manifold structure as Euclidean spaces: a) the geodesic exponential map  $Exp_p : T_p \Sigma_\tau \rightarrow \Sigma_\tau$  is a diffeomorphism ; b) the sectional curvature is less or equal zero everywhere; c) they have no “conjugate locus” [i.e. there are no pairs of conjugate Jacobi points (intersection points of distinct geodesics through them) on any geodesic] and no “cut locus” [i.e. no closed geodesics through any point].

v) Like in Yang-Mills case [48], the 3-spin-connection on the orthogonal frame  $SO(3)$ -bundle (and therefore cotriads) will have to be restricted to suited weighted Sobolev spaces to avoid Gribov ambiguities [48, 78]. In turn, this implies the absence of isometries of the noncompact Riemannian 3-manifold  $(\Sigma_\tau, {}^3g)$  [see for instance the review paper in Ref. [79]].

Diffeomorphisms on  $\Sigma_\tau$  ( $Diff \Sigma_\tau$ ) are interpreted in the passive way, following Ref.[80], in accord with the Hamiltonian point of view that infinitesimal diffeomorphisms are generated by taking the Poisson bracket with the 1st class supermomentum constraints [passive diffeomorphisms are also named ‘pseudodiffeomorphisms’].

The new formulation of tetrad gravity [see Refs. [81] for the existing versions of the theory] utilizes the ADM action of metric gravity with the 4-metric expressed in terms of arbitrary cotriads. Let us remark that both in the ADM metric and tetrad formulation one has to introduce the extra ingredient of the 3+1 splittings of  $M^4$  with foliations whose leaves  $\Sigma_\tau$  are spacelike 3-hypersurfaces. However, their points  $z^\mu(\tau, \vec{\sigma})$  [ $(\tau, \vec{\sigma})$  are  $\Sigma_\tau$ -adapted holonomic coordinates of  $M^4$ ] are not configurational variables of these theories in contrast to what happens in Minkowski parametrized theories as already said [ $\frac{\partial z^\mu}{\partial \sigma^A}$  are not tetrads when  $M^4$  is not Minkowski spacetime with Cartesian coordinates, because  ${}^4g^{AB} \frac{\partial z^\mu}{\partial \sigma^A} \frac{\partial z^\nu}{\partial \sigma^B} = {}^4g^{\mu\nu} \neq {}^4\eta^{(\mu)(\nu)}$ ].

By using  $\Sigma_\tau$ -adapted holonomic coordinates for  $M^4$ , one has found a new parametrization of arbitrary tetrads and cotriads on  $M^4$  in terms of cotriads on  $\Sigma_\tau$  [ ${}^3e_{(a)r}(\tau, \vec{\sigma})$ ], of lapse  $[N(\tau, \vec{\sigma})]$  and shift  $[N_{(a)}(\tau, \vec{\sigma}) = \{ {}^3e_{(a)r} N^r \}(\tau, \vec{\sigma})]$  functions and of 3 parameters  $[\varphi_{(a)}(\tau, \vec{\sigma})]$  parametrizing point-dependent Wigner boosts for timelike Poincaré orbits. Putting these variables in the ADM action for metric gravity [73] (with the 3-metric on  $\Sigma_\tau$  expressed in terms of cotriads:  ${}^3g_{rs} = {}^3e_{(a)r} {}^3e_{(a)s}$  with positive signature), one gets a new action depending only on lapse, shifts and cotriads, but not on the boost parameters (therefore, there is no need to use Schwinger’s time gauge). There are 10 primary and 4 secondary first class constraints and a weakly vanishing canonical Hamiltonian containing the secondary constraints like in ADM metric gravity [73]. Besides the 3 constraints associated with the vanishing Lorentz boost momenta (Abelianization of boosts), there are 4 constraints saying that the momenta associated with lapse and shifts vanish, 3 constraints describing rotations, 3 constraints generating space-diffeomorphisms on the cotriads induced by those ( $Diff \Sigma_\tau$ ) on  $\Sigma_\tau$  (a linear combination of supermomentum constraints and of the rotation ones; a different combination of these constraints generates  $SO(3)$  Gauss law constraints for the momenta  ${}^3\pi_{(a)}$  conjugated to cotriads with the covariant derivative built with the spin connection) and one superhamiltonian constraint. The six constraints connected with Lorentz boosts and rotations replace the constraints satisfying the Lorentz algebra in the older formulations. The boost parameters  $\varphi_{(a)}(\tau, \vec{\sigma})$  and the three angles  $\alpha_{(a)}(\tau, \vec{\sigma})$  hidden in the cotriads are the extra variables of tetrad gravity with respect to metric gravity: they allow a Hamiltonian description of the congruences of timelike accelerated observers used in the formulation of gravitomagnetism[82, 83].

It turns out that with the technology developed for Yang-Mills theory, one can Abelianize the 3 rotation constraints and then also the space-diffeomorphism constraints so that we can arrive at a total of 13 Abelianized first class constraints. In the Abelianization of the rotation constraints one needs the Green function of the 3-dimensional covariant derivative containing the spin connection, well defined only if there is no Gribov ambiguity in the  $SO(3)$ -frame bundle and no isometry of the Riemannian 3-manifold  $(\Sigma_\tau, {}^3g)$ . The Green function is similar to the Yang-Mills one for a principal  $SO(3)$ -bundle [48], but, instead of the Dirac distribution for the Green function of the flat divergence, it contains the Synge-DeWitt bitensor [84] defining the tangent in one endpoint of the geodesic arc connecting two points (which reduces to the Dirac distribution only locally in normal coordinates). Moreover, the definition of the Green function now requires the geodesic exponential

map.

In the resulting quasi-Shanmugadhasan canonical basis, the original cotriad can be expressed in closed form in terms of 3 rotation angles, 3 diffeomorphism-parameters and a reduced cotriad depending only on 3 independent variables (they are Dirac's observables with respect to 13 of the 14 first class constraints) and with their conjugate momenta, still subject to the reduced form of the superhamiltonian constrain: this is the phase space over the superspace of 3-geometries[85].

Till now no coordinate condition[86] has been imposed. It turns out that these conditions are hidden in the choice of how to parametrize the reduced cotriads in terms of three independent functions. The simplest parametrization (the only one studied till now) corresponds to choose a system of global 3-orthogonal coordinates on  $\Sigma_\tau$ , in which the 3-metric is diagonal. With a further canonical transformation on the reduced cotriads and conjugate momenta, one arrives at a canonical basis containing the conformal factor  $\phi(\tau, \vec{\sigma}) = e^{q(\tau, \vec{\sigma})/2}$  of the 3-geometry and its conjugate momentum  $\rho(\tau, \vec{\sigma})$  plus two other pairs of conjugate canonical variables  $r_{\bar{a}}(\tau, \vec{\sigma})$ ,  $\pi_{\bar{a}}(\tau, \vec{\sigma})$ ,  $\bar{a} = 1, 2$ . The reduced superhamiltonian constraint, expressed in terms of these variables, turns out to be an integro-differential equation for the conformal factor (reduced Lichnerowicz equation) whose conjugate momentum is, therefore, the last gauge variable. If we replace the gauge fixing of the Lichnerowicz[87] and York[88, 89, 83] approach [namely the vanishing of the trace of the extrinsic curvature of  $\Sigma_\tau$ ,  ${}^3K(\tau, \vec{\sigma}) \approx 0$ , also named the internal extrinsic York time[90]] with the natural one  $\rho(\tau, \vec{\sigma}) \approx 0$  and we go to Dirac brackets, we find that  $r_{\bar{a}}(\tau, \vec{\sigma})$ ,  $\pi_{\bar{a}}(\tau, \vec{\sigma})$  are the canonical basis for the physical degrees of freedom or Dirac's observables of the gravitational field in the 3-orthogonal gauges. Let us remark that the functional form of the non-tensorial objects  $r_{\bar{a}}$ ,  $\pi_{\bar{a}}$ , depends on the chosen coordinate condition.

The next step is to find the physical Hamiltonian for them and to solve the deparametrization problem. If we wish to arrive at the soldering of tetrad gravity with matter and parametrized Minkowski formulation for the same matter, we must require that the lapse and shift functions of tetrad gravity [which must grow linearly in  $\vec{\sigma}$ , in suitable asymptotic Minkowski coordinates, according to the existing literature on asymptotic Poincaré charges at spatial infinity [74]] must agree asymptotically with the flat lapse and shift functions, which, however, are unambiguously defined only on Minkowski spacelike hyperplanes as we have seen.

In metric ADM gravity the canonical Hamiltonian is  $H_{(c)ADM} = \int d^3\sigma [N\tilde{\mathcal{H}} + N_r\tilde{\mathcal{H}}^r](\tau, \vec{\sigma}) \approx 0$ , where  $\tilde{\mathcal{H}}(\tau, \vec{\sigma}) \approx 0$  and  $\tilde{\mathcal{H}}^r(\tau, \vec{\sigma}) \approx 0$  are the superhamiltonian and supermomentum constraints. It is differentiable and finite only for suitable  $N(\tau, \vec{\sigma}) = n(\tau, \vec{\sigma}) \rightarrow_{|\vec{\sigma}| \rightarrow \infty} 0$ ,  $N_r(\tau, \vec{\sigma}) = n_r(\tau, \vec{\sigma}) \rightarrow_{|\vec{\sigma}| \rightarrow \infty} 0$  defined by Beig and Ó'Murchadha[74] in suitable asymptotic coordinate systems. For more general lapse and shift functions one must add a surface term [85] to  $H_{(c)ADM}$ , which contains the "strong" Poincaré charges [73]  $P_{ADM}^A$ ,  $J_{ADM}^{AB}$  [they are conserved and gauge invariant surface integrals]. To have well defined asymptotic Poincaré charges at spatial infinity[73, 74] one needs: i) the selection of a class of coordinates systems for  $\Sigma_\tau$  asymptotic to flat coordinates; ii) the choice of a class of Hamiltonian boundary conditions for the fields in these coordinate systems [all the fields must belong to some functional space of the type of the weighted Sobolev spaces]; iii) a definition of the Hamiltonian group  $\mathcal{G}$  of gauge transformations (and in particular of proper gauge transformations) with a well defined limit at spatial infinity so to respect i) and ii). The scheme is the same needed to define the non-Abelian charges in Yang-Mills theory[48]. The delicate point is to be able to exclude supertranslations[76], because the presence of these extra asymptotic charges leads to the replacement of the asymptotic Poincaré group with the infinite-dimensional spi group[75] of asymptotic symmetries, which does not allow the definition of the Poincaré spin due to the absence of the Pauli-Lubanski Casimir. This can be done with suitable boundary conditions (in particular all the fields and gauge transformations must have direction independent limits at spatial infinity) respecting the "parity conditions" of Beig and Ó'Murchadha[74].

Let us then remark that in Ref.[91] and in the book in Ref.[1] (see also Ref.[74]), Dirac introduced asymptotic Minkowski rectangular coordinates

$$z_{(\infty)}^{(\mu)}(\tau, \vec{\sigma}) = x_{(\infty)}^{(\mu)}(\tau) + b_{(\infty)}^{(\mu)}{}_r(\tau)\sigma^r$$

in  $M^4$  at spatial infinity  $S_\infty = \cup_\tau S_{\tau,\infty}^2$ . For each value of  $\tau$ , the coordinates  $x_{(\infty)}^{(\mu)}(\tau)$  labels a point, near spatial infinity chosen as origin of  $\Sigma_\tau$ . On it there is a flat tetrad  $b_{(\infty)A}^{(\mu)}(\tau) = \{l_{(\infty)}^{(\mu)} = b_{(\infty)\tau}^{(\mu)} = \epsilon^{(\mu)}_{(\alpha)(\beta)(\gamma)} b_{(\infty)1}^{(\alpha)}(\tau) b_{(\infty)2}^{(\beta)}(\tau) b_{(\infty)3}^{(\gamma)}(\tau); b_{(\infty)r}^{(\mu)}(\tau)\}$ , with  $l_{(\infty)}^{(\mu)}$   $\tau$ -independent, satisfying  $b_{(\infty)A}^{(\mu)} \eta_{(\mu)(\nu)} b_{(\infty)B}^{(\nu)} = \eta_{AB}$  for every  $\tau$  [at this level we do not assume that  $l_{(\infty)}^{(\mu)}$  is tangent to  $S_\infty$ , as the normal  $l^\mu$  to  $\Sigma_\tau$ ]. There will be transformation coefficients  $b_A^\mu(\tau, \vec{\sigma})$  from the holonomic adapted coordinates  $\sigma^A = (\tau, \sigma^r)$  to coordinates  $x^\mu = z^\mu(\sigma^A)$  in an atlas of  $M^4$ , such that in a chart at spatial infinity one has  $z^\mu(\tau, \vec{\sigma}) = \delta_{(\mu)}^\mu z^{(\mu)}(\tau, \vec{\sigma})$  and  $b_A^\mu(\tau, \vec{\sigma}) = \delta_{(\mu)}^\mu b_{(\infty)A}^{(\mu)}(\tau)$  [for  $r \rightarrow \infty$  one has  ${}^4g_{\mu\nu} \rightarrow \delta_{(\mu)}^\mu \delta_{(\nu)}^\nu \eta_{(\mu)(\nu)}$  and  ${}^4g_{AB} = b_A^\mu {}^4g_{\mu\nu} b_B^\nu \rightarrow b_{(\infty)A}^{(\mu)} \eta_{(\mu)(\nu)} b_{(\infty)B}^{(\nu)} = \eta_{AB}$ ].

Dirac[91] and, then, Regge and Teitelboim[74] proposed that the asymptotic Minkowski rectangular coordinates  $z_{(\infty)}^{(\mu)}(\tau, \vec{\sigma}) = x_{(\infty)}^{(\mu)}(\tau) + b_{(\infty)r}^{(\mu)}(\tau) \sigma^r$  should define 10 new independent degrees of freedom at the spatial boundary  $S_\infty$ , as it happens for Minkowski parametrized theories[55] when restricted to spacelike hyperplanes [defined by  $z^{(\mu)}(\tau, \vec{\sigma}) \approx x_s^{(\mu)}(\tau) + b_r^{(\mu)}(\tau) \sigma^r$ ]; then, 10 conjugate momenta should exist. These 20 extra variables of the Dirac proposal can be put in the form:  $x_{(\infty)}^{(\mu)}(\tau)$ ,  $p_{(\infty)}^{(\mu)}$ ,  $b_{(\infty)A}^{(\mu)}(\tau)$  [with  $b_{(\infty)\tau}^{(\mu)} = l_{(\infty)}^{(\mu)}$   $\tau$ -independent],  $S_{(\infty)}^{(\mu)(\nu)}$ , with Dirac brackets implying the orthonormality constraints  $b_{(\infty)A}^{(\mu)} \eta_{(\mu)(\nu)} b_{(\infty)B}^{(\nu)} = \eta_{AB}$  [so that  $p_{(\infty)}^{(\mu)}$  and  $J_{(\infty)}^{(\mu)(\nu)} = x_{(\infty)}^{(\mu)} p_{(\infty)}^{(\nu)} - x_{(\infty)}^{(\nu)} p_{(\infty)}^{(\mu)} + S_{(\infty)}^{(\mu)(\nu)}$  satisfy a Poincaré algebra]. In analogy with Minkowski parametrized theories restricted to spacelike hyperplanes, one expects to have 10 extra first class constraints of the type

$$p_{(\infty)}^{(\mu)} - P_{ADM}^{(\mu)} \approx 0, \quad S_{(\infty)}^{(\mu)(\nu)} - S_{ADM}^{(\mu)(\nu)} \approx 0$$

with  $P_{ADM}^{(\mu)}$ ,  $S_{ADM}^{(\mu)(\nu)}$  related to the ADM Poincaré charges  $P_{ADM}^A$ ,  $J_{ADM}^{AB}$ . The origin  $x_{(\infty)}^{(\mu)}$  is going to play the role of a decoupled observer with his parametrized clock.

Let us remark that if we replace  $p_{(\infty)}^{(\mu)}$  and  $S_{(\infty)}^{(\mu)(\nu)}$ , whose Poisson algebra is the direct sum of an Abelian algebra of translations and of a Lorentz algebra, with the new variables (with holonomic indices with respect to  $\Sigma_\tau$ )  $p_{(\infty)}^A = b_{(\infty)(\mu)}^A p_{(\infty)}^{(\mu)}$ ,  $J_{(\infty)}^{AB} = b_{(\infty)(\mu)}^A b_{(\infty)(\nu)}^B S_{(\infty)}^{(\mu)(\nu)}$  [ $\neq b_{(\infty)(\mu)}^A b_{(\infty)(\nu)}^B J_{(\infty)}^{(\mu)(\nu)}$ ], the Poisson brackets for  $p_{(\infty)}^{(\mu)}$ ,  $b_{(\infty)A}^{(\mu)}$ ,  $S_{(\infty)}^{(\mu)(\nu)}$  imply that  $p_{(\infty)}^A$ ,  $J_{(\infty)}^{AB}$  satisfy a Poincaré algebra. This implies that the Poincaré generators  $P_{ADM}^A$ ,  $J_{ADM}^{AB}$  define in the asymptotic Dirac rectangular coordinates a momentum  $P_{ADM}^{(\mu)}$  and only an ADM spin tensor  $S_{ADM}^{(\mu)(\nu)}$  [to define an angular momentum tensor  $J_{ADM}^{(\mu)(\nu)}$  one should find a “center of mass of the gravitational field”  $X_{ADM}^{(\mu)}[{}^3g, {}^3\tilde{\Pi}]$  (see Ref.[58] for the Klein-Gordon case) conjugate to  $P_{ADM}^{(\mu)}$ , so that  $J_{ADM}^{(\mu)(\nu)} = X_{ADM}^{(\mu)} P_{ADM}^{(\nu)} - X_{ADM}^{(\nu)} P_{ADM}^{(\mu)} + S_{ADM}^{(\mu)(\nu)}$ ].

The following splitting of the lapse and shift functions and the following set of boundary conditions fulfill all the previous requirements [soldering with the lapse and shift functions on Minkowski hyperplanes; absence of supertranslations [strictly speaking one gets  $P_{ADM}^r = 0$  due to the parity conditions;  $r = |\vec{\sigma}|$ ]

$${}^3g_{rs}(\tau, \vec{\sigma}) \rightarrow_{r \rightarrow \infty} (1 + \frac{M}{r}) \delta_{rs} + {}^3h_{rs}(\tau, \vec{\sigma}) = (1 + \frac{M}{r}) \delta_{rs} + o_4(r^{-3/2}),$$

$${}^3\tilde{\Pi}^{rs}(\tau, \vec{\sigma}) \rightarrow_{r \rightarrow \infty} {}^3k^{rs}(\tau, \vec{\sigma}) = o_3(r^{-5/2}),$$

$$N(\tau, \vec{\sigma}) = N_{(as)}(\tau, \vec{\sigma}) + n(\tau, \vec{\sigma}), \quad n(\tau, \vec{\sigma}) = O(r^{-(3+\epsilon)}),$$

$$N_r(\tau, \vec{\sigma}) = N_{(as)r}(\tau, \vec{\sigma}) + n_r(\tau, \vec{\sigma}), \quad n_r(\tau, \vec{\sigma}) = O(r^{-\epsilon}),$$

$$N_{(as)A}(\tau, \vec{\sigma}) \stackrel{def}{=} (N_{(as)}; N_{(as)r})(\tau, \vec{\sigma}) = -\tilde{\lambda}_A(\tau) - \frac{1}{2} \tilde{\lambda}_{As}(\tau) \sigma^s,$$

$$\Rightarrow {}^3e_{(a)r}(\tau, \vec{\sigma}) = (1 + \frac{M}{2r}) \delta_{(a)r} + o_4(r^{-3/2}),$$

with  ${}^3h_{rs}(\tau, -\vec{\sigma}) = {}^3h_{rs}(\tau, \vec{\sigma})$ ,  ${}^3k^{rs}(\tau, -\vec{\sigma}) = -{}^3k^{rs}(\tau, \vec{\sigma})$ ; here  ${}^3\tilde{\Pi}^{rs}(\tau, \vec{\sigma})$  is the momentum conjugate to the 3-metric  ${}^3g_{rs}(\tau, \vec{\sigma})$  in ADM metric gravity.

These boundary conditions identify the class of spacetimes of Christodoulou and Klainermann[92] (they are near to Minkowski spacetime in a norm sense, contain gravitational radiation but evade the singularity theorems, because they do not satisfy the hypothesis of conformal completion to get the possibility to put control on the large time development of the solutions of Einstein's equations). These spacetimes also satisfy the rest-frame condition  $P_{ADM}^r = 0$  (this requires  $\tilde{\lambda}_{Ar}(\tau) = 0$  like for Wigner hyperplanes in parametrized Minkowski theories) and have vanishing shift functions (but non trivial lapse function).

After the addition of the surface term, the resulting canonical and Dirac Hamiltonians of ADM metric gravity are

$$\begin{aligned} H_{(c)ADM} &= \int d^3\sigma [(N_{(as)} + n)\tilde{\mathcal{H}} + (N_{(as)r} + n_r){}^3\tilde{\mathcal{H}}^r](\tau, \vec{\sigma}) \mapsto \\ &\mapsto H'_{(c)ADM} = \int d^3\sigma [(N_{(as)} + n)\tilde{\mathcal{H}} + (N_{(as)r} + n_r){}^3\tilde{\mathcal{H}}^r](\tau, \vec{\sigma}) + \\ &+ \tilde{\lambda}_A(\tau)P_{ADM}^A + \tilde{\lambda}_{AB}(\tau)J_{ADM}^{AB} = \\ &= \int d^3\sigma [n\tilde{\mathcal{H}} + n_r{}^3\tilde{\mathcal{H}}^r](\tau, \vec{\sigma}) + \tilde{\lambda}_A(\tau)\hat{P}_{ADM}^A + \tilde{\lambda}_{AB}(\tau)\hat{J}_{ADM}^{AB} \approx \\ &\approx \tilde{\lambda}_A(\tau)\hat{P}_{ADM}^A + \tilde{\lambda}_{AB}(\tau)\hat{J}_{ADM}^{AB}, \end{aligned}$$

with the “weak conserved improper charges”  $\hat{P}_{ADM}^A$ ,  $\hat{J}_{ADM}^{AB}$  [they are volume integrals differing from the weak charges by terms proportional to integrals of the constraints]. The previous splitting implies to replace the variables  $N(\tau, \vec{\sigma})$ ,  $N_r(\tau, \vec{\sigma})$  with the ones  $\tilde{\lambda}_A(\tau)$ ,  $\tilde{\lambda}_{AB}(\tau) = -\tilde{\lambda}_{BA}(\tau)$ ,  $n(\tau, \vec{\sigma})$ ,  $n_r(\tau, \vec{\sigma})$  [with conjugate momenta  $\tilde{\pi}^A(\tau)$ ,  $\tilde{\pi}^{AB}(\tau) = -\tilde{\pi}^{BA}(\tau)$ ,  $\tilde{\pi}^n(\tau, \vec{\sigma})$ ,  $\tilde{\pi}_n^r(\tau, \vec{\sigma})$ ] in the ADM theory.

With these assumptions one has the following form of the line element (also its form in tetrad gravity is given)

$$\begin{aligned} ds^2 &= \epsilon([N_{(as)} + n]^2 - [N_{(as)r} + n_r]^3 e_{(a)}^r e_{(a)}^s [N_{(as)s} + n_s])(d\tau)^2 - \\ &- 2\epsilon[N_{(as)r} + n_r]d\tau d\sigma^r - \epsilon^3 e_{(a)r}^3 e_{(a)s}^3 d\sigma^r d\sigma^s. \end{aligned}$$

The final suggestion of Dirac is to modify ADM metric gravity in the following way:

- i) add the 10 new primary constraints  $p_{(\infty)}^A - \hat{P}_{ADM}^A \approx 0$ ,  $J_{(\infty)}^{AB} - \hat{J}_{ADM}^{AB} \approx 0$ , where  $p_{(\infty)}^A = b_{(\infty)(\mu)}^A p_{(\infty)}^{(\mu)}$ ,  $J_{(\infty)}^{AB} = b_{(\infty)(\mu)}^A b_{(\infty)(\nu)}^B S_{(\infty)}^{(\mu)(\nu)}$  [remember that  $p_{(\infty)}^A$  and  $J_{(\infty)}^{AB}$  satisfy a Poincaré algebra];
- ii) consider  $\tilde{\lambda}_A(\tau)$ ,  $\tilde{\lambda}_{AB}(\tau)$ , as Dirac multipliers for these 10 new primary constraints, and not as configurational (arbitrary gauge) variables coming from the lapse and shift functions [so that there are no conjugate (vanishing) momenta  $\tilde{\pi}^A(\tau)$ ,  $\tilde{\pi}^{AB}(\tau)$  and no associated Dirac multipliers  $\zeta_A(\tau)$ ,  $\zeta_{AB}(\tau)$ ], in the assumed Dirac Hamiltonian [it is finite and differentiable]

$$\begin{aligned} H_{(D)ADM} &= \int d^3\sigma [n\tilde{\mathcal{H}} + n_r\tilde{\mathcal{H}}^r + \lambda_n\tilde{\pi}^n + \lambda_r^{\tilde{\pi}}\tilde{\pi}_n^r](\tau, \vec{\sigma}) - \\ &- \tilde{\lambda}_A(\tau)[p_{(\infty)}^A - \hat{P}_{ADM}^A] - \tilde{\lambda}_{AB}(\tau)[J_{(\infty)}^{AB} - \hat{J}_{ADM}^{AB}] \approx 0, \end{aligned}$$

The reduced phase space is still the ADM one: on the ADM variables there are only the secondary first class constraints  $\tilde{\mathcal{H}}(\tau, \vec{\sigma}) \approx 0$ ,  $\tilde{\mathcal{H}}^r(\tau, \vec{\sigma}) \approx 0$  [generators of proper gauge transformations], because the other first class constraints  $p_{(\infty)}^A - \hat{P}_{ADM}^A \approx 0$ ,  $J_{(\infty)}^{AB} - \hat{J}_{ADM}^{AB} \approx 0$  do not generate improper gauge transformations but eliminate 10 of the extra 20 variables.

In this modified ADM metric gravity, one has restricted the 3+1 splittings of  $M^4$  to foliations whose leaves  $\Sigma_\tau$  tend to Minkowski spacelike hyperplanes asymptotically at spatial infinity in a direction independent way. Therefore, these  $\Sigma'_\tau$  should be determined by the 10 degrees of freedom  $x_{(\infty)}^{(\mu)}(\tau)$ ,  $b_{(\infty)A}^{(\mu)}(\tau)$ , like it happens for flat spacelike hyperplanes: this means that it must be possible to define a “parallel transport” of the asymptotic tetrads  $b_{(\infty)A}^{(\mu)}(\tau)$  to get well defined tetrads in each point of  $\Sigma'_\tau$ . While it is not yet clear whether this can be

done for  $\tilde{\lambda}_{AB}(\tau) \neq 0$ , there is a solution for  $\tilde{\lambda}_{AB}(\tau) = 0$ . This case corresponds to go to the Wigner-like hypersurfaces [the analogue of the Minkowski Wigner hyperplanes with the asymptotic normal  $l_{(\infty)}^{(\mu)} = l_{(\infty)\Sigma}^{(\mu)}$  parallel to  $\hat{P}_{ADM}^{(\mu)}$ ]. Following the same procedure defined for Minkowski spacetime, one gets  $\tilde{S}_{(\infty)}^{rs} \equiv \hat{J}_{ADM}^{rs}$  [see Ref.[55] for the definition of  $\tilde{S}_{(\infty)}^{AB}$ ],  $\tilde{\lambda}_{AB}(\tau) = 0$  and  $-\tilde{\lambda}_A(\tau)[p_{(\infty)}^A - \hat{P}_{ADM}^A] = -\tilde{\lambda}_\tau(\tau)[\epsilon_{(\infty)} - \hat{P}_{ADM}^\tau] + \tilde{\lambda}_\tau(\tau)\hat{P}_{ADM}^\tau$  [ $\epsilon_{(\infty)} = \sqrt{p_{(\infty)}^2}$ ], so that the final form of these four surviving constraints is ( $P_{ADM}^\tau = 0$  implies  $\hat{P}_{ADM}^\tau \approx 0$ ;  $M_{ADM} = \sqrt{\hat{P}_{ADM}^2} \approx \hat{P}_{ADM}^\tau$  is the ADM mass of the universe)

$$\epsilon_{(\infty)} - \hat{P}_{ADM}^\tau \approx 0, \quad \hat{P}_{ADM}^\tau \approx 0.$$

On this subclass of foliations [whose leaves  $\Sigma_\tau^{(WSW)}$  will be called Wigner-Sen-Witten hypersurfaces; they define the intrinsic asymptotic rest frame of the gravitational field] one can introduce a parallel transport by using the interpretation of Ref.[93] of the Witten spinorial method of demonstrating the positivity of the ADM energy [94]. Let us consider the Sen-Witten connection [95, 94] restricted to  $\Sigma_\tau^{(WSW)}$  (it depends on the trace of the extrinsic curvature of  $\Sigma_\tau^{(WSW)}$ ) and the spinorial Sen-Witten equation associated with it. As shown in Ref.[96], this spinorial equation can be rephrased as an equation whose solution determines (in a surface dependent dynamical way) a tetrad in each point of  $\Sigma_\tau^{(WSW)}$  once it is given at spatial infinity (again this requires a direction independent limit). Therefore, at spatial infinity there is a privileged congruence of time-like observers, which replaces the concept of “fixed stars” in the study of the precessional effects of gravitomagnetism on gyroscopes and whose connection with the definition of post-Newtonian coordinates has still to be explored.

On the Wigner-Sen-Witten hypersurfaces the spatial indices have become spin-1 Wigner indices [they transform with Wigner rotations under asymptotic Lorentz transformations]. As said for parametrized theories in Minkowski spacetime, in this special gauge 3 degrees of freedom of the gravitational field [an internal 3-center-of-mass variable  $\vec{\sigma}_{ADM}$  [ ${}^3g, {}^3\vec{\Pi}$ ] inside the Wigner-Sen-Witten hypersurface] become gauge variables, while  $\tilde{x}_{(\infty)}^{(\mu)}$  [the canonical non covariant variable replacing  $x_{(\infty)}^{(\mu)}$ ] becomes a decoupled observer with his “point particle clock” [4, 5] near spatial infinity. Since the positivity theorems for the ADM energy imply that one has only timelike or lightlike orbits of the asymptotic Poincaré group, the restriction to universes with timelike ADM 4-momentum allows to define the Møller radius  $\rho_{AMD} = \sqrt{-\hat{W}_{ADM}^2/\hat{P}_{ADM}^2}$  from the asymptotic Poincaré Casimirs  $\hat{P}_{ADM}^2, \hat{W}_{ADM}^2$ .

By going from  $\tilde{x}_{(\infty)}^{(\mu)}, p_{(\infty)}^{(\mu)}$ , to the canonical basis  $T_{(\infty)} = p_{(\infty)(\mu)}\tilde{x}_{(\infty)}^{(\mu)}/\epsilon_{(\infty)} = p_{(\infty)(\mu)}x_{(\infty)}^{(\mu)}/\epsilon_{(\infty)}$ ,  $\epsilon_{(\infty)}, z_{(\infty)}^{(i)} = \epsilon_{(\infty)}(\tilde{x}_{(\infty)}^{(i)} - p_{(\infty)}^{(i)}\tilde{x}_{(\infty)}^{(o)}/p_{(\infty)}^{(o)})$ ,  $k_{(\infty)}^{(i)} = p_{(\infty)}^{(i)}/\epsilon_{(\infty)} = u^{(i)}(p_{(\infty)}^{(\rho)})$ , like in the flat case one finds that the final reduction requires the gauge-fixings  $T_{(\infty)} - \tau \approx 0$  and  $\sigma_{ADM}^\tau \approx 0$ , where  $\sigma^\tau = \sigma_{ADM}^\tau$  is a variable representing the “internal center of mass” of the 3-metric of the slice  $\Sigma_\tau$  of the asymptotically flat spacetime  $M^4$ . Since  $\{T_{(\infty)}, \epsilon_{(\infty)}\} = -\epsilon$ , with the gauge fixing  $T_{(\infty)} - \tau \approx 0$  one gets  $\tilde{\lambda}_\tau(\tau) \approx \epsilon$ , and the final Dirac Hamiltonian is  $H_D = M_{ADM} + \tilde{\lambda}_\tau(\tau)\hat{P}_{ADM}^\tau$  with  $M_{ADM}$  the natural physical Hamiltonian to reintroduce an evolution in the “mathematical”  $T_{(\infty)} \equiv \tau$ : namely in the rest-frame time identified with the parameter  $\tau$  labelling the leaves  $\Sigma_\tau^{(WSW)}$  of the foliation of  $M^4$ . Physical times (atomic clocks, ephemeris time...) must be put in a local 1-1 correspondence with this “mathematical” time. This point of view excludes any Wheeler-DeWitt interpretation of an internal time (like the extrinsic York one or the WKB times), which is used in closed universes of the Einstein-Wheeler type.

All this construction holds also in our formulation of tetrad gravity (since it uses the ADM action) and in its canonically reduced form in the 3-orthogonal gauges. The final physical Hamiltonian of tetrad gravity for the physical gravitational field is the reduced volume form of the ADM energy  $\hat{P}_{ADM}^\tau[r_{\vec{a}}, \pi_{\vec{a}}, \phi(r_{\vec{a}}, \pi_{\vec{a}})]$  with the conformal factor  $\phi$  solution of the reduced Lichnerowicz equation in the 3-orthogonal gauge with  $\rho(\tau, \vec{\sigma}) \approx 0$ . The Hamilton-Dirac equations generated



by this Hamiltonian for  $\tau_{\bar{a}}$ ,  $\pi_{\bar{a}}$  generate the pair of second order equations in normal form for  $\tau_{\bar{a}}$  hidden in the Einstein equations in this particular gauge.

Let us compare the standard generally covariant formulation of gravity based on the Hilbert action with its invariance under  $Diff M^4$  with the ADM Hamiltonian formulation.

Regarding the 10 Einstein equations of the standard approach, the Bianchi identities imply that four equations are linearly dependent on the other six ones and their gradients. Moreover, the four combinations of Einstein's equations projectable to phase space (where they become the secondary first class superhamiltonian and supermomentum constraints of canonical metric gravity) are independent from the accelerations being restrictions on the Cauchy data. As a consequence the Einstein equations have solutions, in which the ten components  ${}^4g_{\mu\nu}$  of the 4-metric depend on only two truly dynamical degrees of freedom (defining the physical gravitational field) and on eight undetermined degrees of freedom. This transition from the ten components  ${}^4g_{\mu\nu}$  of the tensor  ${}^4g$  in some atlas of  $M^4$  to the 2 (deterministic)+8 (undetermined) degrees of freedom breaks general covariance, because these quantities are neither tensors nor invariants under diffeomorphisms (their functional form is atlas dependent).

Since the Hilbert action is invariant under  $Diff M^4$ , one usually says that a "dynamical gravitational field" is a 4-geometry over  $M^4$ , namely an equivalence class of spacetimes  $(M^4, {}^4g)$ , solution of Einstein's equations, modulo  $Diff M^4$ . See, however, the interpretational problems about what is observable in general relativity for instance in Refs.[7, 8], in particular the facts that at least before the restriction to the solutions of Einstein's equations i) scalars under  $Diff M^4$ , like  ${}^4R$ , are not Dirac's observables but gauge dependent quantities; ii) the functional form of  ${}^4g_{\mu\nu}$  in terms of the physical gravitational field and, therefore, the angle and distance properties of material bodies and the standard procedures of defining measures of length and time based on the line element  $ds^2$ , are gauge dependent.

Instead in the ADM formalism with the extra notion of 3+1 splittings of  $M^4$ , the (tetrad) metric ADM action (differing from the Hilbert one by a surface term) is quasi-invariant under the (14) 8 types of gauge transformations which are the pull-back of the Hamiltonian group  $\mathcal{G}$  of gauge transformations, whose generators are the first class constraints of the theory. The Hamiltonian group  $\mathcal{G}$  has a subgroup (whose generators are the supermomentum and superhamiltonian constraints) formed by the diffeomorphisms of  $M^4$  adapted to its 3+1 splittings,  $Diff M^{3+1}$  [it is different from  $Diff M^4$ ]. Moreover, the Poisson algebra of the supermomentum and superhamiltonian constraints reflects the embeddability in  $M^4$  of the foliation associated with the 3+1 splitting [97].

Now in tetrad gravity the interpretation of the 14 gauge transformations and of their gauge fixings (it is independent from the presence of matter) is the following [a tetrad in a point of  $\Sigma_\tau$  is a local observer] :

- i) the gauge fixings of the gauge boost parameters associated with the 3 boost constraints and of the gauge angles associated with the 3 rotation constraints are equivalent to choose the congruence of timelike observers to be used as a standard of non rotation;
- ii) the gauge fixings of the 3 gauge parameters associated with the passive space diffeomorphisms [ $Diff \Sigma_\tau$ ; change of coordinates charts] are equivalent to a fixation of 3 standards of length by means of a choice of a coordinate system on  $\Sigma_\tau$  [the measuring apparatus (the "rods") should be defined in terms of Dirac's observables for some kind of matter, after its introduction into the theory];
- iii) according to constraint theory the choice of 3-coordinates on  $\Sigma_\tau$  induces the gauge fixings of the 3 shift functions [i.e. of  ${}^4g_{0i}$ ], whose gauge nature is connected with the "conventionality of simultaneity" [98] [therefore, the gauge fixings are equivalent to a choice of synchronizaton of clocks and, as a consequence, to a statement about the isotropy or anisotropy of the velocity of light in that gauge];
- iv) the gauge fixing on the the momentum  $\rho(\tau, \vec{\sigma})$  conjugate to the conformal factor of the 3-metric [this gauge variable is the source of the gauge dependence of 4-tensors and of the scalars under  $Diff M^4$ , together with the gradients of the lapse and shift functions] is a nonlocal statement about the extrinsic curvature of the leaves  $\Sigma_\tau$  of the given 3+1 splitting of  $M^4$ ; since the superhamiltonian constraint produces normal deformations of  $\Sigma_\tau$  [97] and, therefore, transforms a 3+1 splitting of



$M^4$  into another one (the ADM formulation is independent from the choice of the 3+1 splitting), this gauge fixing is equivalent to the choice of a particular 3+1 splitting;  
 v) the previous gauge fixing induces the gauge fixing of the lapse function (which determines the packing of the leaves  $\Sigma_\tau$  in the chosen 3+1 splitting) and, therefore, is equivalent to the fixation of a standard of proper time [again “clocks” should be built with the Dirac’s observables of some kind of matter].

In the Hamiltonian formalism it is natural to define a “Hamiltonian kinematical gravitational field” as the equivalence class of spacetimes modulo the Hamiltonian group  $\mathcal{G}$ , and different members of the equivalence class have in general different 4-Riemann tensors [these equivalence classes are connected with the conformal 3-geometries of the Lichnerowicz-York approach and contain different gauge-related 4-geometries]. Then, a “Hamiltonian dynamical gravitational field” is defined as a Hamiltonian kinematical gravitational fields which is solution of the Hamilton-Dirac equations generated by the weak ADM energy  $\hat{P}_{ADM}^\tau$ . Since the Hilbert and ADM actions, even if they have different local symmetries and invariances, both generate the same Einstein equations, the equivalence classes of the “Hamiltonian dynamical gravitational fields” and of the standard “dynamical gravitational fields” (a 4-geometry solution of Einstein’s equations) coincide. Indeed, on the solutions of Einstein’s equations the gauge transformations generated by the superhamiltonian constraint (normal deformations of  $\Sigma_\tau$ ) and those generated by the canonical momenta of the lapse and shift functions together with the  $\Sigma_\tau$  diffeomorphisms generated by the supermomentum constraints are restricted by the Jacobi equations associated to Einstein’s equations to be those Noether symmetries of the ADM action which are also dynamical symmetries of the Hamilton equations and therefore they are a subset of the spacetime diffeomorphisms  $Diff M^4$  (all of which are dynamical symmetries of Einstein’s equations).

The 3-orthogonal gauges of tetrad gravity are the equivalent of the Coulomb gauge in classical electrodynamics (like the harmonic gauge is the equivalent of the Lorentz gauge). Only after a complete gauge fixing the 4-tensors and the scalars under  $Diff M^4$  become measurable quantities (like the electromagnetic vector potential in the Coulomb gauge): an experimental laboratory does correspond by definition to a completely fixed gauge. At this stage it becomes acceptable the proposal of Komar[99] and Bergmann[80] of identifying the points of a spacetime  $(M^4, {}^4g)$ , solution of the Einstein’s equations in absence of matter, in a way invariant under spacetime diffeomorphisms, by using four bilinears and trilinears in the Weyl tensors, scalar under  $Diff M^4$  and called “individuating fields” (see also Refs.[7, 8]), which do not depend on the lapse and shift functions (but only on the gauge variables corresponding to the 3-coordinates on  $\Sigma_\tau$  and to the momentum conjugate to the conformal factor of the 3-metric, so that these fields carry the information on the choice of the 3-coordinates and of a generalized extrinsic time), to build “physical 4-coordinates” (in each completely fixed gauge they depend only on the two canonical pairs of Dirac’s observables of the gravitational field), justifying a posteriori the standard measurement theory presented in all textbooks on general relativity, which presupposes the individuation of spacetime points.

Our approach breaks the general covariance of general relativity completely by going to the special 3-orthogonal gauges. But this is done in a way naturally associated with theories with first class constraints: the global Shanmugadhasan canonical transformations (when they exist) correspond to privileged Darboux charts for presymplectic manifolds defined by the first class constraints. Therefore, the gauges identified by these canonical transformations should have a special (till now unexplored) role also in generally covariant theories, in which traditionally one looks for observables invariant under diffeomorphisms and not for not generally covariant Dirac observables.

Let us remember that Bergmann[80] made the following critique of general covariance: it would be desirable to restrict the group of coordinate transformations (spacetime diffeomorphisms) in such a way that it could contain an invariant subgroup describing the coordinate transformations that change the frame of reference of an outside observer (these transformations could be called Lorentz transformations; see also the comments in Ref.[100] on the asymptotic behaviour of coordinate transformations); the remaining coordinate transformations would be like the gauge transformations of electromagnetism. This is what we have done. In this way “preferred” coordinate systems will emerge (the WSW hypersurfaces with their preferred congruences of timelike

observers whose 4-velocity becomes asymptotically normal to  $\Sigma_\tau^{(WSW)}$  at spatial infinity), which, as said by Bergmann, are not “flat”: while the inertial coordinates are determined experimentally by the observation of trajectories of force-free bodies, these intrinsic coordinates can be determined only by much more elaborate experiments (probably with gyroscopes), since they depend, at least, on the inhomogeneities of the ambient gravitational fields. See also Ref.[101] for other critics to general covariance: very often to get physical results one uses preferred coordinates not merely for calculational convenience, but also for understanding (this fact has been formalized as the “principle of restricted covariance”).

Since in the 3-orthogonal gauges we have the physical canonical basis  $r_{\bar{a}}, \pi_{\bar{a}}$ , it is possible, but only in absence of matter, to define “void spacetimes” as the equivalence class of spacetimes “without gravitational field”, whose members in the 3-orthogonal gauges are obtained by adding by hand the second class constraints  $r_{\bar{a}}(\tau, \vec{\sigma}) \approx 0, \pi_{\bar{a}}(\tau, \vec{\sigma}) \approx 0$  [one gets  $\phi(\tau, \vec{\sigma}) = 1$  as the relevant solution of the reduced Lichnerowicz equation] and, in particular, their Poincaré charges vanish (this corresponds to the exceptional  $p^{(\mu)} = 0$  orbit of the Poincaré group and shows the peculiarity of these solutions with zero ADM mass). It is expected that the void spacetimes can be defined in a gauge-independent way by adding to the ADM action the requirement that the leaves  $\Sigma_\tau$  of the 3+1 splitting be 3-conformally flat, namely that the Cotton-York 3-conformal tensor vanishes. The members of this equivalence class (the extension to general relativity of the Galilean non inertial coordinate systems with their Newtonian inertial forces) are gauge equivalent to Minkowski spacetime with Cartesian coordinates and it is expected that they describe pure acceleration effects without physical gravitational field (no tidal effects).

See Ref.[102] for the  $c \rightarrow \infty$  contraction of the ADM action of metric gravity: a theory with 26 independent fields (most of them describe inertial forces) and with general Galileo covariance has been obtained. This formulation of Newton gravity should be the natural nonrelativistic limit of Einstein’s general relativity in the framework of singular Lagrangians; however, its connection with the post-Newtonian approximations has still to be explored.

If we add [72] to the tetrad ADM action the action for  $N$  scalar particles with positive energy in the form of Ref.[55] [where it was given on arbitrary Minkowski spacelike hypersurfaces], the only constraints which are modified are the superhamiltonian one, which gets a dependence on the matter energy density  $\mathcal{M}(\tau, \vec{\sigma})$ , and the 3 space diffeomorphism ones, which get a dependence on the matter momentum density  $\mathcal{M}_\tau(\tau, \vec{\sigma})$ . The canonical reduction and the determination of the Dirac observables can be done like in absence of matter. However, the reduced Lichnerowicz equation for the conformal factor of the 3-metric in the 3-orthogonal gauge and with  $\rho(\tau, \vec{\sigma}) \approx 0$  acquires now an extra dependence on  $\mathcal{M}(\tau, \vec{\sigma})$  and  $\mathcal{M}_\tau(\tau, \vec{\sigma})$ .

Since, as a preliminary result, we are interested in identifying explicitly the instantaneous action-at-a-distance (Newton-like and gravitomagnetic) potentials among particles hidden in tetrad gravity (like the Coulomb potential is hidden in the electromagnetic gauge potential), we shall make the strong approximation of neglecting the (tidal) effects of the physical gravitational field by putting  $r_{\bar{a}}(\tau, \vec{\sigma}) \approx 0, \pi_{\bar{a}}(\tau, \vec{\sigma}) \approx 0$ , even if it is not strictly consistent with the Hamilton-Dirac equation (extremely weak gravitational fields). If, furthermore, we develop the conformal factor  $\phi(\tau, \vec{\sigma})$  in a formal series in the Newton constant  $G$  [ $\phi = 1 + \sum_{n=1}^{\infty} G^n \phi_n$ ], one can find a solution  $\phi = 1 + G\phi_1$  at order  $G$  (post-Minkowskian approximation) of the reduced Lichnerowicz equation where we put  $r_{\bar{a}} = \pi_{\bar{a}} = 0$ . However, due to a self-energy divergence in  $\phi$  evaluated at the positions  $\vec{\eta}_i(\tau)$  of the particles, one needs to rescale the bare masses to physical ones,  $m_i \mapsto \phi^{-2}(\tau, \vec{\eta}_i(\tau))m_i^{(phys)}$ , and to make a regularization of the type defined in Refs. [103]. Then, the regularized solution for  $\phi$  can be put in the reduced form of the ADM energy, which becomes [ $\vec{\kappa}_i(\tau)$  are the particle momenta conjugate to  $\vec{\eta}_i(\tau)$ ;  $\vec{n}_{ij} = [\vec{\eta}_i - \vec{\eta}_j]/|\vec{\eta}_i - \vec{\eta}_j|$ ]

$$\begin{aligned} \hat{P}_{ADM}^\tau = & \sum_{i=1}^N c \sqrt{m_i^{(phys)2} c^2 + \vec{\kappa}_i^2(\tau)} - \\ & - \frac{G}{c^2} \sum_{i \neq j} \frac{\sqrt{m_i^{(phys)2} c^2 + \vec{\kappa}_i^2(\tau)} \sqrt{m_j^{(phys)2} c^2 + \vec{\kappa}_j^2(\tau)}}{|\vec{\eta}_i(\tau) - \vec{\eta}_j(\tau)|} - \\ & - \frac{G}{8c^2} \sum_{i \neq j} \frac{3\vec{\kappa}_i(\tau) \cdot \vec{\kappa}_j(\tau) - 5\vec{\kappa}_i(\tau) \cdot \vec{n}_{ij}(\tau) \vec{\kappa}_j(\tau) \cdot \vec{n}_{ij}(\tau)}{|\vec{\eta}_i(\tau) - \vec{\eta}_j(\tau)|} + O(G^2, r_{\bar{a}}, \pi_{\bar{a}}). \end{aligned}$$

One sees the Newton-like and the gravitomagnetic (in the sense of York) potentials (both of them need regularization) at the post-Minkowskian level (order  $G$  but exact in  $c$ ) emerging from the tetrad ADM version of Einstein general relativity when we ignore the tidal effects. For  $G=0$  we recover  $N$  free scalar particles on the Wigner hyperplane in Minkowski spacetime, as required by deparametrization. For  $c \rightarrow \infty$ , we get the post-Newtonian Hamiltonian

$$H_{PN} = \sum_{i=1}^N \frac{\tilde{\kappa}_i^2(\tau)}{2m_i^{(phys)}} \left(1 - \frac{2G}{c^2} \sum_{j \neq i} \frac{m_j^{(phys)}}{|\tilde{\eta}_i(\tau) - \tilde{\eta}_j(\tau)|}\right) - \frac{G}{2} \sum_{i \neq j} \frac{m_i^{(phys)} m_j^{(phys)}}{|\tilde{\eta}_i(\tau) - \tilde{\eta}_j(\tau)|} - \frac{G}{8c^2} \sum_{i \neq j} \frac{3\tilde{\kappa}_i(\tau) \cdot \tilde{\kappa}_j(\tau) - 5\tilde{\kappa}_i(\tau) \cdot \tilde{n}_{ij}(\tau) \tilde{\kappa}_j(\tau) \cdot \tilde{n}_{ij}(\tau)}{|\tilde{\eta}_i(\tau) - \tilde{\eta}_j(\tau)|} + O(G^2, r_a, \pi_a),$$

which is of the type of the ones implied by the results of Refs.[103, 104] [the differences are probably connected with the use of different coordinate systems and with the fact that one has essential singularities on the particle worldlines and the need of regularization].

The main open problems now under investigation are: i) the linearization of the theory in the 3-orthogonal gauges in presence of matter to find the 3-orthogonal gauge description of gravitational waves and to go beyond the previous instantaneous post-Minkowskian approximation at least in the 2-body case relevant for the motion of binaries; ii) the replacement of scalar particles with spinning ones to identify the precessional effects (like the Lense-Thirring one) of gravitomagnetism; iii) the coupling to perfect fluids for the simulation of rotating stars and for the comparison with the post-Newtonian approximations; iv) the coupling of tetrad gravity to the electromagnetic field, to fermion fields and then to the standard model, trying to make to reduction to Dirac's observables in all these cases and to study their post-Minkowskian approximations; v) the quantization of tetrad gravity in the 3-orthogonal gauge with  $\rho(\tau, \vec{\sigma}) \approx 0$  (namely after a complete breaking of general covariance): for each perturbative (in  $G$ ) solution of the reduced Lichnerowicz equation one defines a Schrodinger equation in  $\tau$  for a wave functional  $\Psi[\tau; r_a]$  with the associated quantized ADM energy  $\hat{P}_{ADM}^\tau[r_a, i\frac{\delta}{\delta r_a}]$  as Hamiltonian; no problem of physical scalar product is present, but only ordering problems in the Hamiltonian; moreover, one has the Møller radius as a ultraviolet cutoff. Also a comparison with "loop quantum gravity" [105], which respects general covariance but only for fixed lapse and shift functions, has still to be done.

Therefore, a well defined classical stage for a unified description of the four interactions is emerging, even if many aspects have only been clarified at a heuristic level so that a big effort from both mathematical and theoretical physicists is still needed. It will be exciting to see whether in the next years some reasonable quantization picture will develop from this classical framework.

## References

- [1] P.A.M.Dirac, *Can.J.Math.* **2**, 129 (1950); "Lectures on Quantum Mechanics", Belfer Graduate School of Science, Monographs Series (Yeshiva University, New York, N.Y., 1964).
- [2] J.L.Anderson and P.G.Bergmann, *Phys.Rev.* **83**, 1018 (1951).  
P.G.Bergmann and J.Goldberg, *Phys.Rev.* **98**, 531 (1955).
- [3] S.Weinberg, "The Theory of Fields", 2 volumes (Cambridge Univ.Press, Cambridge, 1995 and 1996).
- [4] C.J.Isham, "Canonical Quantum Gravity and the Problem of Time", in "Integrable Systems, Quantum Groups and Quantum Field Theories", eds.L.A.Ibort and M.A.Rodriguez, Salamanca 1993 (Kluwer, London, 1993); "Conceptual and Geometrical Problems in Quantum Gravity", in "Recent Aspects of Quantum Fields", Schladinger 1991, eds. H.Mitter and H.Gausterer (Springer, Berlin, 1991); "Prima Facie Questions in Quantum Gravity" and "Canonical Quantum Gravity and the Question of Time", in "Canonical Gravity: From Classical to Quantum", eds. J.Ehlers and H.Friedrich (Springer, Berlin, 1994).

- [5] K.Kuchar, "Time and Interpretations of Quantum Gravity", in Proc.4th Canadian Conf. on "General Relativity and Relativistic Astrophysics", eds. G.Kunstatter, D.Vincent and J.Williams (World Scientific, Singapore, 1992).
- [6] J.Butterfield and C.J.Isham, "Space-Time and the Philosophical Challenge of Quantum Gravity", Imperial-TP-98-99-45 (gr-qc/9903072).
- [7] J.Stachel, in "General Relativity and Gravitation", GR11, Stockholm 1986, ed. M.A.H.Mac Callum (Cambridge Univ. Press, Cambridge, 1987); "The Meaning of General Covariance", in "Philosophical Problems of the Internal and External Worlds", Essays in the Philosophy of A.Grünbaum, eds. J.Earman, A.I.Janis, G.J.Massey and N.Rescher (Pittsburgh Univ.Press, Pittsburgh, 1993).
- [8] C.Rovelli, *Class.Quantum Grav.* **8**, 297 and 317 (1991).
- [9] J.Polchinski, "String Theory", 2 volumes (Cambridge Univ. Press, Cambridge, 1998).
- [10] L.Lusanna, "Solving Gauss' Laws and Searching Dirac Observables for the Four Interactions", talk at the "Second Conf. on Constrained Dynamics and Quantum Gravity", S.Margherita Ligure 1996, eds. V.De Alfaro, J.E.Nelson, G.Bandelloni, A.Biasi, M.Cavaglià and A.T.Filippov, *Nucl.Phys. (Proc.Suppl.)* **B57**, 13 (1997) (HEP-TH/9702114). "Unified Description and Canonical Reduction to Dirac's Observables of the Four Interactions", talk at the Int.Workshop "New non Perturbative Methods and Quantization on the Light Cone", Les Houches School 1997, eds. P.Grangé, H.C.Pauli, A.Neveu, S.Pinsky and A.Werner (Springer, Berlin, 1998) (HEP-TH/9705154). "The Pseudoclassical Relativistic Quark Model in the Rest-Frame Wigner-Covariant Gauge", talk at the Euroconference QCD97, ed. S.Narison, Montpellier 1997, *Nucl.Phys. (Proc. Suppl.)* **B64**, 306 (1998).
- [11] T.Levi-Civita, *Prace Mat.Fiz.* **17**, 1 (1906); T.Levi-Civita and U.Amaldi, "Lezioni di Meccanica Razionale", Vol.II, Part 2 (Zanichelli, Bologna, 1927).
- [12] A.Lichnerowicz, *C.R.Acad.Sci.Paris, Ser. A*, **280**, 523 (1975). W.Tulczyiew, *Symposia Math.* **14**, 247 (1974). N.Woodhouse, "Geometric Quantization" (Clarendon, Oxford, 1980). J.Śniatycki, *Ann.Inst. H.Poincaré* **20**, 365 (1984). G.Marmo, N.Mukunda and J.Samuel, *Riv.Nuovo Cimento* **6**, 1 (1983). M.J.Bergvelt and E.A.De Kerf, *Physica* **139A**, 101 and 125 (1986). B.A.Dubrovinn, M.Giordano, G.Marmo and A.Simoni, *Int.J.Mod.Phys.* **8**, 4055 (1993).
- [13] M.J.Gotay, J.M.Nester and G.Hinds, *J.Math.Phys.* **19**, 2388 (1978). M.J.Gotay and J.M.Nester, *Ann.Inst.Henri Poincaré* **A30**, 129 (1979) and **A32**, 1 (1980). M.J.Gotay and J.Śniatycki, *Commun. Math.Phys.* **82**, 377 (1981). M.J.Gotay, *Proc.Am.Math.Soc.* **84**, 111 (1982); *J.Math.Phys.* **27**, 2051 (1986).
- [14] M.Henneaux and C.Teitelboim, "Quantization of Gauge Systems" (Princeton Univ. Press, Princeton, 1992).
- [15] R.Sugano, Y.Kagraoka and T.Kimura, *Int.J.Mod.Phys.* **7**, 61 (1992).
- [16] J.Śniatycki, in "Non-Linear Partial Differential Operators and Quantization Procedures", Clausthal 1981, *Lecture Notes Math.* 1037 (Springer, Berlin, 1983). J.Śniatycki and A.Weinstein, *Lett.Math.Phys.* **7**, 155 (1983).
- [17] P.G.Bergmann, *Phys.Rev.* **144**, 1078 (1966). K.Kuchar, *J.Math.Phys.* **13**, 758 (1972). K.Komar, *Phys.Rev.* **D18**, 1881, 1887 and 3017 (1978); **D19**, 2908 (1979). D.Dominici, J.Gomis, G.Longhi and J.M.Pons, *J.Math.Phys.* **25**, 2439 (1984).
- [18] L.Lusanna, *Phys.Rep.* **185**, 1 (1990).
- [19] L.Lusanna, *Riv. Nuovo Cimento* **14**, n.3, 1 (1991).

- [20] L.Lusanna, *J.Math.Phys.* **31**, 2126 (1990).
- [21] L.Lusanna, *J.Math.Phys.* **31**, 428 (1990).
- [22] L.Lusanna, *Int.J.Mod.Phys. A* **8**, 4193 (1993).
- [23] M.Chaichian, D.Louis Martinez and L.Lusanna, *Ann.Phys.(N.Y.)* **232**, 40 (1994).
- [24] E.Noether, *Nachr.Ges.Wiss.Göttingen, Math.Phys.Kl.H.* **2**, 235 (1918); English translation in *Transp.Theory Stat.Phys.* **1**, 183 (1971). J.D.Logan, "Invariant Variational Principles" (Academic, New York, 1979). N.P.Konopleva and V.N.Popov, "Gauge Fields" (Harwood, New York, 1981). B.M.Barbashov and V.V.Nesterenko, *Fortschr.Phys.* **31**, 535 (1983).
- [25] I.A.Batalin and G.A.Vilkoviski, *Nucl.Phys.* **B234**, 106 (1984).
- [26] S.Shanmugadhasan, *J.Math.Phys.* **14**, 677 (1973).
- [27] J.A.Schouten and W.V.D.Kulk, "Pfaff's Problem and Its Generalizations" (Clarendon, Oxford, 1949).
- [28] S.Lie, "Theorie der Transformation Gruppe", Vol. II (B.G.Teubner, Leipzig, 1890). A.R.Forsyth, "Theory of Differential Equations", Vol. V, Ch. IX (Dover, New York, 1959). L.P.Eisenhart, "Continuous Groups of Transformations (Dover, New York, 1961). R.O.Fulp and J.A.Marlin, *Pacific J. Math.* **67**, 373 (1976); *Rep.Math.Phys.* **18**, 295 (1980).
- [29] L.D.Faddeev and Popov, *Phys.Lett.* **B25**, 30 (1967).
- [30] M.Henneaux, *Phys.Rep.* **126**, 1 (1985).
- [31] J.M.Souriau, "Structure des systémes dynamiques" (Dunod, Paris, 1970). B.Konatant, "Quantization and Unitary Representations", *Lecture Notes Math.* **170** (Springer, Berlin, 1970). J.E.Marsden and A.Weinstein, *Rep.Math.Phys.* **5**, 121 (1974).
- [32] G.Longhi and L.Lusanna, *Phys.Rev.* **D34**, 3707 (1986).
- [33] A.Lucenti, L.Lusanna and M.Pauri, *J.Phys.* **A31**, 1633 (1998).
- [34] J.M.Arms, J.E.Marsden and V.Moncrief, *Commun.Math.Phys.* **78**, 455 (1981). J.M.Arms, *Acta Phys.Pol.* **B17**, 499 (1986). L.Bos and M.J.Gotay, *J.Diff.Geom.* **24**, 181 (1986).
- [35] P.A.M.Dirac, *Can.J.Phys.* **33**, 650 (1955).
- [36] R.Casalbuoni, *Nuovo Cimento* **33A**, 115 and 389 (1976). F.A. Berezin and M.S.Marinov, *Ann.Phys.(N.Y.)* **104**, 336 (1977). A.Barucci, R.Casalbuoni and L.Lusanna, *Nuovo Cim.Lett.* **19**, 581 (1977); *Nucl.Phys.* **B124**, 93 (1981) and **B180**[FS2], 141 (1981).
- [37] A.Barducci, R.Casalbuoni and L.Lusanna, *Nuovo Cim.* **35A**, 377 (1976).
- [38] A.Barducci, R.Casalbuoni, D.Dominici and L.Lusanna, *Phys.Lett.* **100B**, 126 (1981).
- [39] A.Barducci and L.Lusanna, *Nuovo Cim.* **77A**, 39 (1983).
- [40] A.Barducci and L.Lusanna, *J.Phys.* **A16**, 1993 (1983).
- [41] Ph.Droz Vincent, *Lett.Nuovo Cim.* **1**, 839 (1969); *Phys.Scr.* **2**, 129 (1970); *Rep.Math.Phys.* **8**, 79 (1975). I.T.Todorov, Report Comm. JINR E2-10125, Dubna 1976 (unpublished); *Ann.Inst.H.Poincaré* **28A**, 207 (1978). A.Komar, *Phys.Rev.* **D18**, 1881 and 1887 (1978).
- [42] A.Barducci, R.Casalbuoni and L.Lusanna, *Nuovo Cim.* **54A**, 340 (1979).

- [43] H.Sazdjian, *Ann.Phys.(N.Y.)* **136**, 136 (1981); *Phys.Rev.* **D33**, 3401 (1986); *J.Math.Phys.* **28**, 2618 and 1988 (1987), **29**, 1620 (1987); *Ann.Phys. (N.Y.)* **191**, 52 (1989); in *Proc.Int. Symp. "Extended Objects and Bound Systems"*, eds. O.Hara, S.Ishida and S.Naka (World Scientific, Singapore, 1992). J.Bijtebier and J.Brockaert, *Nuovo Cim.* **A105**, 351 and 625 (1992); in *Proc.Int.Symp. "Extended Objects and Bound Systems"*, eds. O.Hara, S.Ishida and S.Naka (World Scientific, Singapore, 1992).
- [44] H.W.Crater and P.Van Alstine, *J.Math.Phys.* **23**, 1697 (1982); *Ann.Phys.(N.Y.)* **148**, 57 (1983); *Phys.Rev.Lett.* **53**, 1577 (1984); *Phys.Rev.* **D30**, 2585 (1984); **D34**, 1932 (1986); **D36**, 3007 (1987); **D37**, 1982 (1988); *J.Math.Phys.* **31**, 1998 (1990); *Phys.Rev.* **D46**, 766 (1992). H.W.Crater, R.L. Becker, C.Y.Wong and P.Van Alstine, *Phys.Rev.* **D46**, 5117 (1992); in *Proc.Int.Symp. "Extended Objects and Bound Systems"*, eds. O.Hara, S.Ishida and S.Naka (World Scientific, Singapore, 1992). H.W.Crater and D.Yang, *J.Math.Phys.* **32**, 2374 (1991).
- [45] L.Lusanna, *Nuovo Cim.* **64A**, 65 (1981).
- [46] G.Longhi, L.Lusanna and J.M.Pons, *J.Math.Phys.* **30**, 1893 (1989).
- [47] F.Colomo, G.Longhi and L.Lusanna, *Int.J.Mod.Phys.* **A5**, 3347 (1990); *Mod.Phys.Letters* **A5**, 17 (1990). F.Colomo and L.Lusanna, *Int.J.Mod.Phys.* **A7**, 1705 and 4107 (1992).
- [48] L.Lusanna, *Int.J.Mod.Phys.* **A10**, 3531 and 3675 (1995).
- [49] H.Cendra, A.Ibort and J.Marsden, *J.Geom.Phys.* **4**, 183 (1987). A.P.Balachandran, G.Marmo, B.S.Skagerstam and A.Stern, "Classical Topology and Quantum States" (World Scientific, Singapore, 1991).
- [50] V.Moncrief, *J.Math.Phys.* **20**, 579 (1979). M.Cantor, *Bull.Am.Math.Soc.* **5**, 235 (1981).
- [51] L.Lusanna and P.Valtancoli, *Int.J.Mod.Phys.* **A12**, 4769 (1997) (HEP-TH/9606078) and *Int.J.Mod.Phys.* **A12**, 4797 (1997). (HEP-TH/9606079).
- [52] L.Lusanna and P.Valtancoli, *Int.J.Mod.Phys.* **A13**, 4605 (1998) (HEP-TH/9707072).
- [53] K.Kuchar, *J.Math.Phys.* **17**, 777, 792, 801 (1976); **18**, 1589 (1977).
- [54] P.A.M.Dirac, *Rev.Mod.Phys.* **21** (1949) 392.
- [55] L.Lusanna, *Int.J.Mod.Phys.* **A12**, 645 (1997).
- [56] M.Pauri and M.Prosperi, *J.Math.Phys.* **16**, 1503 (1975).
- [57] L.Lusanna and M.Materassi, "The Canonical Decomposition in Collective and Relative Variables of a Klein-Gordon Field in the Rest-Frame Wigner-Covariant Instant Form", Firenze Univ.preprint (HEP-TH/9904202).
- [58] G.Longhi and M.Materassi, *J.Math.Phys.* **40**, 480 (1999) (HEP-TH/9803128); "Collective and Relative Variables for a Classical Klein-Gordon Field", Firenze Univ.preprint (HEP-TH/9890024), to appear in *Int.J.Mod.Phys.* **A**.
- [59] W.G.Dixon, *J.Math.Phys.* **8**, 1591 (1967). "Extended Objects in General Relativity: their Description and Motion", in "Isolated Gravitating Systems in General Relativity", ed.J.Ehlers (North-Holland, Amsterdam, 1979).
- [60] D.Alba and L.Lusanna, *Int.J.Mod.Phys.* **A13**, 2791 (1998) (HEP-TH/9705155).
- [61] D.Alba and L.Lusanna, *Int.J.Mod.Phys.* **A13**, 3275 (1998) (HEP-TH/9705156).
- [62] F.Bigazzi and L.Lusanna, *Int.J.Mod.Phys.* **A14**, 1429 (1999) (HEP-TH/9807052).
- [63] F.Bigazzi and L.Lusanna, *Int.J.Mod.Phys.* **A14**, 1877 (1999) (HEP-TH/9807054).

- [64] C.Lämmerzahl, *J.Math.Phys.* **34**, 3918 (1993).
- [65] I.Herbst, *Commun.Math.Phys.* **53**, 285 (1977); **55**, 316 (1997).  
 B. and L. Durand, *Phys.Rev.* **D28**, 396 (1983); erratum *Phys.Rev.* **D50**, 6642 (1994).  
 J.J.Basdevant and S.Boukraa, *Z.Phys.* **C28**, 413 (1985).  
 A.Martin and S.M.Roy, *Phys.Lett.* **B233**, 407 (1989).  
 A.LeYaouanc, L.Oliver and J.C.Raynal, *Ann.Phys.(N.Y.)* **239**, 243 (1995).  
 W.Lucha and F.F.Schöberl, *Phys.Rev.* **D50**, 5443 (1994).
- [66] G.Leibbrandt, "Non-Covariant Gauges", ch.9 (World Scientific, Singapore, 1994).
- [67] C.Møller, *Ann.Inst.H.Poincaré* **11**, 251 (1949); "The Theory of Relativity" (Oxford Univ.Press, Oxford, 1957).
- [68] G.Veneziano, "Quantum Strings and the Constants of Nature", in "The Challenging Questions", ed.A.Zichichi, the Subnuclear Series n.27 (Plenum Press, New York, 1990).
- [69] L.Lusanna and S.Russo, "Tetrad Gravity I): A New Formulation", Firenze Univ. preprint 1998 (GR-QC/9807073).
- [70] L.Lusanna and S.Russo, "Tetrad Gravity II): Dirac's Observables", Firenze Univ. preprint 1998 (GR-QC/9807074).
- [71] R.DePietri and L.Lusanna, "Tetrad Gravity III): Asymptotic Poincaré Charges, the Physical Hamiltonian and Void Spacetimes", in preparation.
- [72] R.DePietri, L.Lusanna and M.Vallisneri, "Tetrad Gravity IV): The N-body Problem", in preparation.
- [73] R.Arnowitt, S.Deser and C.W.Misner, *Phys.Rev.* **117**, 1595 (1960); in "Gravitation: an Introduction to Current Research", ed.L.Witten (Wiley, New York, 1962).
- [74] T.Regge and C.Teitelboim, *Ann.Phys.(N.Y.)* **88**, 286 (1974). R.Beig and Ó Murchadha, *Ann.Phys.(N.Y.)* **174**, 463 (1987). L.Andersson, *J.Gem.Phys.* **4**, 289 (1987).
- [75] A.Ashtekar, "Asymptotic Structure of the Gravitational Field at Spatial Infinity", in "General Relativity and Gravitation", Vol. 2, ed.A.Held (Plenum, New York, 1980). A.Ashtekar and R.O.Hansen, *J.Math.Phys.* **19**, 1542 (1978). A.Ashtekar and A.Magnon, *J.Math.Phys.* **25**, 2682 (1984). A.Ashtekar and J.D.Romano, *Class.Quantum Grav.* **9**, 1069 (1992).
- [76] R.M.Wald, "General Relativity" (Chicago Univ.Press, Chicago, 1984).
- [77] R.Geroch, *J.Math.Phys.* **9**, 1739 (1968); **11**, 343 (1970).
- [78] V.Moncrief, *J.Math.Phys.* **16**, 1556 (1975).
- [79] Y.Choquet-Bruhat, A.Fischer and J.E.Marsden, "Maximal Hypersurfaces and Positivity of Mass", LXVII E.Fermi Summer School of Physics "Isolated Gravitating Systems in General Relativity", ed.J.Ehlers (North-Holland, Amsterdam, 1979).
- [80] P.G.Bergmann, *Rev.Mod.Phys.* **33**, 510 (1961).
- [81] H.Weyl, *Z.Physik* **56**, 330 (1929). J.Schwinger, *Phys.Rev.* **130**, 1253 (1963). T.W.B.Kibble, *J.Math.Phys.* **4**, 1433 (1963). S.Deser and C.J.Isham, *Phys.Rev.* **D14**, 2505 (1976). J.E.Nelson and C.Teitelboim, *Ann.Phys.(N.Y.)* **116**, 86 (1978). M.Pilati, *Nucl.Phys.* **B132**, 138 (1978). J.E.Nelson and T.Regge, *Ann.Phys.(N.Y.)* **166**, 234 (1986); *Int.J.Mod.Phys.* **A4**, 2021 (1989). J.M.Charap and J.E.Nelson, *J.Phys.* **A16**, 1661 and 3355 (1983). *Class.Quantum Grav.* **3**, 1061 (1986).  
 J.M.Charap, "The Constraints in Vierbein General Relativity", in "Constraint's Theory and Relativistic Dynamics", eds. G.Longhi and L.Lusanna (World Scientific, Singapore,

- 1987). M.Henneaux, *Gen.Rel.Grav.* **9**, 1031 (1978). M.Henneaux, *Phys.Rev.* **D27**, 986 (1983). J.M.Charap, M.Henneaux and J.E.Nelson, *Class.Quantum Grav.* **5**, 1405 (1988). M.Henneaux, J.E.Nelson and C.Schonblond, *Phys.Rev.* **D39**, 434 (1989).
- [82] R.T.Jantzen, P.Carini and D.Bini, *Ann.Phys.(N.Y.)* **215**, 1 (1992).
- [83] I.Ciufolini and J.A.Wheeler, "Gravitation and Inertia" (Princeton Univ.Press, Princeton, 1995).
- [84] J.L.Synge, "Relativity: the General Theory" (North-Holland, Amsterdam, 1960). B.S.De Witt, *Phys.Rev.* **162**, 1195 (1967); "The Spacetime Approach to Quantum Field Theory", in "Relativity, Groups and Topology II", Les Houches 1983, eds. B.S.DeWitt and R.Stora (North-Holland, Amsterdam, 1984).
- [85] B.S.De Witt, *Phys.Rev.* **160**, 1113 (1967).
- [86] C.J.Isham and K.Kuchar, *Ann.Phys.(N.Y.)* **164**, 288 and 316 (1984). K.Kuchar, *Found.Phys.* **16**, 193 (1986).
- [87] A.Lichnerowicz, *J.Math.Pure Appl.* **23**, 37 (1944). Y.Choquet-Bruhat, *C.R.Acad.Sci.Paris* **226**, 1071 (1948); *J.Rat.Mech.Anal.* **5**, 951 (1956); "The Cauchy Problem" in "Gravitation: An Introduction to Current Research", ed.L.Witten (Wiley, New York, 1962).
- [88] J.W.York jr, *Phys.Rev.Lett.* **26**, 1656 (1971); **28**, 1082 (1972). *J.Math.Phys.* **13**, 125 (1972); **14**, 456 (1972). *Ann.Inst.H.Poincaré* **XXI**, 318 (1974). N.O'Murchadha and J.W.York jr, *J.Math.Phys.* **14**, 1551 (1972). *Phys.Rev.* **D10**, 428 (1974).
- [89] J.W.York jr., "Kinematics and Dynamics of General Relativity", in "Sources of Gravitational Radiation", Battelle-Seattle Workshop 1978, ed.L.L.Smarr (Cambridge Univ.Press, Cambridge, 1979).
- [90] A.Qadir and J.A.Wheeler, "York's Cosmic Time Versus Proper Time", in "From SU(3) to Gravity", Y.Ne'eman's festschrift, eds. E.Gotsma and G.Tauber (Cambridge Univ.Press, Cambridge, 1985).
- [91] P.A.M.Dirac, *Canad.J.Math.* **3**, 1 (1951).
- [92] D.Christodoulou and S.Klainerman, "The Global Nonlinear Stability of the Minkowski Space" (Princeton Univ. Press, Princeton, 1993).
- [93] A.Ashtekar and G.T.Horowitz, *J.Math.Phys.* **25**, 1473 (1984).
- [94] E.Witten, *Commun.Math.Phys.* **80**, 381 (1981).
- [95] A.Sen, *J.Math.Phys.* **22**, 1781 (1981); *Phys.Lett.* **119B**, 89 (1982).
- [96] J.Frauenfelder, *Class.Quantum Grav.* **8**, 1881 (1991).
- [97] C.Teitelboim, "The Hamiltonian Structure of Space-Time", in "General Relativity and Gravitation", ed.A.Held, Vol.I (Plenum, New York, 1980). A.S.Hojman, K.Kuchar and C.Teitelboim, *Ann.Phys. (N.Y.)* **96**, 88 (1971).
- [98] P.Havas, *Gen.Rel.Grav.* **19**, 435 (1987). R.Anderson, I.Vetharaniam and G.E.Stedman, *Phys.Rep.* **295**, 93 (1998).
- [99] A.Komar, *Phys.Rev.* **111**, 1182 (1958). P.G.Bergmann and A.Komar, *Phys.Rev.Lett.* **4**, 432 (1960).
- [100] L.Landau and E.Lifschitz, "The Classical Theory of Fields" (Addison-Wesley, Cambridge, 1951).



- [101] G.F.R.Ellis and D.R.Matracers, *Gen.Rel.Grav.* **27**, 777 (1995). R.Zalaletdinov, R.Tavakol and G.F.R.Ellis, *Gen.Rel.Grav.* **28**, 1251 (1996).
- [102] R.DePietri, L.Lusanna and M.Pauri, *Class.Quantum Grav*, **12**, 219 (1995).
- [103] A.Einstein, B.Hoffman and L.Infeld, *Ann.Math.* **39**, 66 (1938). A.Einstein and L.Infeld, *Ann.Math.* **41**, 797 (1940); *Canad.J.Math.* **1**, 209 (1949). L.Infeld, *Rev.Mod.Phys.* **29**, 398 (1957).
- [104] H.A.Lorentz and L.Droste, *Amst.Akad.Versl.* **26**, 392 (1917). A.Eddington and G.L.Clarke, *Proc.Roy.Soc.London* **A166**, 465 (1938). V.Fock, *J.Phys. (U.S.S.R.)* **1**, 81 (1939). A.Papapetrou, *Proc.Phys.Soc.(London)* **64**, 57 (1951).
- [105] A.Ashtekar, *Phys.Rev.Lett.* **57**, 2244 (1986); "New Perspectives in Canonical Gravity" (Bibliopolis, Naples, 1988); "Lectures on Non-Perturbative Canonical Gravity" (World Scientific, Singapore, 1991); "Quantum Mechanics of Riemannian Geometry", [http://vishnu.nirvana.phys.psu.edu/riem\\_qm/riem\\_qm.html](http://vishnu.nirvana.phys.psu.edu/riem_qm/riem_qm.html). C.Rovelli and L.Smolín, *Nucl.Phys.* **B331**, 80 (1990); **B442**, 593 (1995). C.Rovelli, "Loop Quantum Gravity", *Living Reviews in Relativity* <http://www.livingreviews.org/Articles/Volume1/1998-1rovelli>.

## Part D : Extension Of QFT Frontiers

- 20. Supersymmetry And Particle Physics by R.N.Mohapatra
- 21. Supersymmetry In Field Theory by N.Sakai
- 22. Conformal Field Theory: A Bridge Over Troubled Waters by Werner Nahm
- 23. Superstring Theory - An Overview by John H Schwarz
- 24. Recent Developments In String Theory by J.Maharana
- 25. Yang-Mills Theory And Matrix String Theory by L.Bonora



# 20. Supersymmetry and Particle Physics

R. N. Mohapatra \*

Department of Physics, University of Maryland,  
College Park, MD, 20742, USA

## Abstract

A pedagogical overview of the recent developments in the area of supersymmetry and its applications to particle physics is presented.

## 1 Introduction

All elementary particles in nature fall into two classes: bosons and fermions. The bosons have integral spin whereas the fermions have half-odd integral spin. The different statistics obeyed by the bosons and fermions provides a fundamental distinction between them. It was due to this basic dissimilarity that physicists until the mid seventies thought it impossible to have a symmetry which could link the two fundamental species of particles. Thus it was feared that no matter how unified the ultimate theory of matter and forces becomes, the bosons and fermions would for ever remain distinct.

It therefore caused great exhilaration and anticipation when theorists in the mid seventies announced that local field theories with supersymmetry, a symmetry that unites the fermions and bosons together, can not only be written down but they exhibit much more interesting properties than the existing field theories without such symmetries. It is the purpose of this article to provide a pedagogical overview of some of these astounding developments in supersymmetric field theories and their applications in resolving several important issues in particle physics.

Supersymmetry was introduced in the early 1970's in the two dimensional field theories by Gervais and Sakita and in the context of four dimensional ones by Golfand, Likhtman, Akulov, Volkov, Wess and Zumino and has become one of the most active areas of research in the eighties and nineties. As the millenium comes to a close, in addition to the obvious fact that it provides the hope of an unified understanding of the two known forms of matter, the bosons and fermions, the versatility of supersymmetry as a tool to understand many unsolved problems of physics seems to be continuously expanding. Some of these results that will be touched in this review are (i) the significant improvement in the singularity structure of local field theories and the application of this result to understanding the disparate scales of nature such as the Fermi scale characterising weak interactions and the Planck scale characterising the gravitational interactions; (ii) possibility of unifying gravity with the rest of the forces of nature by making supersymmetry into a local rather than a global symmetry; (iii) finally the novel prospect of understanding nonperturbative properties of field theories, that was considered a practical impossibility in nonsupersymmetric field theories.

Let us quickly hasten to remind the reader that the “traders of supersymmetric merchandise” are fully cognisant of the fact that there is no trace of even an iota of supersymmetry in observed spectrum of particles. This fact instead of being discouraging is accepted as a challenge that in itself might more clearly elucidate the nature of supersymmetry and its various links with other

---

\*Email: rmohapat@katherine.physics.umd.edu.

known aspects of physics such as gravity or weak forces. We will therefore spend a full section or two on the various ways that supersymmetry breaking is supposed to emerge. This is a rapidly growing field and our discussion of this aspect will therefore necessarily reflect the tentativeness of this field.

This review is organized as follows: in section 2, the introductory ideas of supersymmetric field theories are presented; in section 3, techniques for writing down the simple supersymmetric field theories are given; section 4 deals with some properties of  $N=1$  super-field theories; section 5, considers the possibility of local supersymmetry and how that brings gravity into the same picture with other interactions of nature, providing the hope for an ultimate unification of all forces including gravity; section 6 discusses breaking of supersymmetry; section 7 reviews some current speculations on the origin of supersymmetry breaking; in section 8, we discuss the application of supersymmetry to particle physics models- i.e. minimal supersymmetric extension of the standard model (MSSM); section 9 discusses the need to go beyond the MSSM; section 10 discusses, the supersymmetric left-right model which cures several of the problems of MSSM and provides a way to include neutrino masses into particle physics; section 11 present the motivation for supersymmetry based on the hypothesis of grand unification of all forces of nature; section 12 discusses possible applications of strongly coupled super-Yang-Mills theories to build composite models of quarks and leptons; section 13 notes the connection between string theory and supersymmetry and section 14 gives the conclusions.

This article is meant to be a pedagogical overview of the field; therefore there are not too many details; many of the discussions are very telegraphic so as to only convey the spirit of the activity in the field and the reader should take this article more as an “appetizer” rather than a “full course dinner” and I will consider my efforts immensely rewarded if I manage to whet a few appetites. There are many excellent books and reviews for those who want to go deeper.

## 2 Superspace and superfields

There are many ways to introduce supersymmetric field theories and we refer to several texts [1, 2] and reviews [3, 4] for more thorough discussion. Here we will content ourselves with a brief outline noting a few salient points.

Since supersymmetry transforms a boson to a fermion and vice versa, an irreducible representation of the supersymmetry “group” contains in it both fermions and bosons. Therefore in a supersymmetric theory, all known particles are accompanied by a superpartner which is a fermion if the known particle is a boson and vice versa. For instance, the electron ( $e$ ) supermultiplet will contain its superpartner  $\tilde{e}$ , (called the selectron) which has spin zero. The photon ( $\gamma$ ) supermultiplet will contain its superpartner,  $\tilde{\gamma}$ , known as the photino, which is a Majorana field of spin  $\frac{1}{2}$ . We will adopt the notation that the superpartner of a particle will be denoted by the same symbol as the particle with a ‘tilde’ as above. The name of the superpartner of any known fermion will have an extra letter ‘s’ at the beginning, and the names of superpartners of known bosons will end with the letters ‘-ino’. Furthermore, while supersymmetry does not commute with the Lorentz transformations, it commutes with all internal symmetries; as a result, all non-Lorentzian quantum numbers for both the fermion and boson in the same supermultiplet are the same. For example, all particles in a supersymmetry multiplet have the same electric charge, same isospin, same color quantum numbers etc.

A convenient mathematical way to discuss supersymmetry is to extend space-time to a superspace where one augments the normal spacetime with the addition of anticommuting Grassmanian coordinates which due to Lorentz covariance can be chosen to form a Majorana spinor. Such spinors have four independent real components or two independent complex components. In what follows we will use the notation of complex two component spinors denoted by  $\theta$  and its complex conjugate  $\bar{\theta}$ . Thus we have exactly the same number of fermionic (or grassmanian) co-ordinates as the number of space time co-ordinates. Models based on this extended space time are called  $N=1$  supersymmetric theories. From this nomenclature it is clear that there can be higher (or extended) supersymmetric theories where for each space-time co-ordinate there are larger number

of fermionic variables.

The two 2-component spinors transform as two dimensional representations of the  $SL(2, C)$  group. This observation has the hidden implication that any function of  $\theta$  and  $\bar{\theta}$  will maintain Lorentz covariance (and invariance) of the field theories constructed out of them as long as we use operators that have proper transformation under the  $SL(2, C)$  group<sup>1</sup>.

Operator valued functions in the superspace will be called superfields and denoted by  $\Phi(x, \theta, \bar{\theta})$ . The Grassmanian property of the  $\theta$  coordinates (i.e.,  $\theta_1\theta_2 = -\theta_2\theta_1$  and  $(\theta_i)^2 = 0$ ) then implies that an expansion of the superfields  $\Phi$  in the  $\theta$  coordinate terminates after a few terms. The coefficients of such an expansion will be functions of  $x$  only, and are therefore fields of the ordinary field theory. Those fields will create and destroy the known particles and their superpartners. The superfields can of course have any spin but for the description of the particles of the standard model, it is enough to consider only those superfields with spin zero.

To see the detailed field theory content of the general superfields, let us expand a general superfield  $\Phi(x, \theta, \bar{\theta})$  in the grassmanian variables:

$$\begin{aligned} \Phi(x, \theta, \bar{\theta}) = & \phi + \theta\psi + \bar{\theta}\bar{\chi} + \theta^2 M + \bar{\theta}^2 N + \theta\sigma^\mu\bar{\theta}V_\mu \\ & \theta^2\bar{\theta}\bar{\lambda} + \bar{\theta}^2\theta\zeta + \theta^2\bar{\theta}^2 D \end{aligned} \quad (1)$$

where  $\sigma^\mu$  are the usual Pauli matrices for the space components and unit matrix for the zeroth component. Note that the fermionic and bosonic fields in the above equation are complex. As a result, simple counting enables one to see that there are 16 real bosonic and same number of real fermionic fields. This equality in the number of the bosonic and fermionic fields will be a constant feature of supersymmetric field theories and play a crucial role in the improved divergence structure of the supersymmetric field theories alluded to above.

The set of fields in Eq. (1) provide a representation multiplet of  $N=1$  supersymmetry. We will see below however that this is not an irreducible representation. To see this let us define the operation of supersymmetry transformation. *Supersymmetry is a translation in the fermionic part of the superspace* i.e. under supersymmetry,  $\theta \rightarrow \theta + \epsilon$  and similarly for  $\bar{\theta}$  where  $\epsilon$  is a constant spinor for the case of global supersymmetry (and a function of space time for the case of local supersymmetry). A key ingredient of supersymmetry transformation is that under supersymmetry transformation described above, the bosonic space coordinates transform as  $x^\mu \rightarrow i(\theta\sigma^\mu\bar{\epsilon} - \epsilon\sigma^\mu\bar{\theta})$ . It is then easy to verify that the generators of supersymmetry are:

$$Q = \frac{\partial}{\partial\theta} - i\sigma^\mu\bar{\theta}\partial_\mu \quad (2)$$

and similarly for  $\bar{Q}$ . The supersymmetry algebra can therefore be written as

$$\{Q, \bar{Q}\} = \sigma^\mu P_\mu \quad (3)$$

In order to construct irreducible representations of supersymmetry, we have to find operators that commute with  $Q$  and  $\bar{Q}$ . There is a pair of such operators called  $\mathcal{D}$  and  $\bar{\mathcal{D}}$  where  $\mathcal{D} = \frac{\partial}{\partial\theta} + i\sigma^\mu\bar{\theta}\partial_\mu$  and similarly for  $\bar{\mathcal{D}}$ . One may therefore impose constraints on the representations such as the  $\Phi$  described above by demanding that  $\mathcal{D}\Phi = 0$  or  $\bar{\mathcal{D}}\Phi = 0$ . These constraints relate the different component fields in the expansion of  $\Phi$  discussed above. One may also put constraints such as  $D^2\Phi = 0$  etc. to get other irreducible representations.

Using the above operators, we can deduce three kinds of spin zero superfields which form irreducible representations of supersymmetry: (i) the chiral and anti-chiral superfields, which result when it is required that either  $\mathcal{D}$  or  $\bar{\mathcal{D}}$  annihilate the field  $\Phi$  (ii) vector superfields which result when it is demanded that  $\Phi = \Phi^\dagger$ ; and (iii) finally linear multiplets result when one requires that  $\mathcal{D}^2$  or  $\bar{\mathcal{D}}^2$  annihilate the  $\Phi$  field. Again for our purpose, we only need the first two kinds of multiplets. The chiral (and anti-chiral) multiplets contain a spin zero and a spin half particle and

<sup>1</sup>Since we visualize the  $\theta$  and  $\bar{\theta}$  as extra "space" variables, we can define a "calculus" on these variables. This is done by defining the  $\int d\theta = 0$ ;  $\int \theta d\theta = 1$  and similarly for the  $\bar{\theta}$ . Differentiation is as usual except for the anticommuting nature of the differential operator.

will be used to describe matter as well as Higgs fields. On the other hand the vector multiplets contain a spin one and a spin half fields and will be used to describe gauge fields. It turns out that real vector fields allow transformations, which can be identified with the gauge transformations as is needed if they are to describe gauge fields.

For completeness, let us introduce the gauge transformations of chiral and the vector (gauge) fields for the abelian example:

$$\begin{aligned}\Phi &\rightarrow e^{-\Lambda}\Phi \\ V &\rightarrow V + \Lambda + \Lambda^\dagger\end{aligned}\quad (4)$$

Note that  $\Lambda$  is a chiral superfield and therefore has not only the familiar gauge parameter of QED which is a real function of space-time, but also a fermionic field and another real field ( $\text{Im } C(x)$  where  $C(x)$  is the scalar component of the superfield  $\Lambda$ ). These extra functions can be used to set the scalar,  $\theta$ ,  $\bar{\theta}$ ,  $\theta^2$  and  $\bar{\theta}^2$  components of  $V$  to zero. This is called the Wess-Zumino gauge, in which the vector field looks like

$$V = \theta\sigma^\mu\bar{\theta}A_\mu + \bar{\theta}^2\theta\lambda + \theta^2\bar{\theta}\bar{\lambda} + \theta^2\bar{\theta}^2 D \quad (5)$$

### 3 The Lagrangian for supersymmetric field theories

In order to write down the action for a supersymmetric field theory, let us start by considering generic chiral fields denoted by  $\Phi(x, \theta)$  with component fields given by  $(\phi, \psi)$  and gauge fields denoted by  $V(x, \theta, \bar{\theta})$  with component gauge and gaugino fields given by  $(A^\mu, \lambda)$ . The action in the superfield notation is

$$S = \int d^4x \int d^2\theta d^2\bar{\theta} \Phi^\dagger e^V \Phi + \int d^4x \int d^2\theta [W(\Phi) + W^\alpha(V)W_\alpha(V) + h.c.] \quad (6)$$

In the above equation, the first term gives the gauge invariant kinetic energy term for the matter fields  $\Phi$ ;  $W(\Phi)$  is a holomorphic function of  $\Phi$  and is called the superpotential; it leads to the Higgs potential of the usual gauge field theories. We wish to emphasize the holomorphy of the superpotential, which is a major difference from the potentials in nonsupersymmetric field theories. This fact has a number of interesting implications that we discuss below.

Secondly,  $W^\alpha(V) \equiv \bar{D}^2 D^\alpha V$  where  $\mathcal{D} \equiv \partial_\theta + i\sigma_\theta \cdot \partial_x$  as defined above, and the term involving  $W^\alpha(V)$  leads to the gauge invariant kinetic energy term for the gauge fields as well as for the gaugino fields. In terms of the component fields the lagrangian can be written as

$$\mathcal{L} = \mathcal{L}_g + \mathcal{L}_{matter} + \mathcal{L}_Y - V(\phi) \quad (7)$$

where

$$\begin{aligned}\mathcal{L}_g &= -\frac{1}{4}F^{\mu\nu}F_{\mu\nu} + \frac{1}{2}\bar{\lambda}\gamma^\mu iD_\mu\lambda \\ \mathcal{L}_{matter} &= |D_\mu\phi|^2 + \bar{\psi}\gamma^\mu iD_\mu\psi \\ \mathcal{L}_Y &= \sqrt{2}g\bar{\lambda}\psi\phi^\dagger + \psi_a\psi_b W_{ab} \\ V(\phi) &= |W_a|^2 + \frac{1}{2}\mathcal{D}_\alpha\mathcal{D}_\alpha\end{aligned}\quad (8)$$

where  $D_\mu$  stands for the covariant derivative with respect to the gauge group and  $\mathcal{D}_\alpha$  stands for the so-called  $\mathcal{D}$ -term and is given by  $\mathcal{D}_\alpha = g\phi^\dagger T_\alpha \phi$  ( $g$  is the gauge coupling constant and  $T_\alpha$  are the generators of the gauge group).  $W_a$  and  $W_{ab}$  are the first and second derivative of the superpotential  $W$  with respect to the superfield with respect to the field  $\Phi_a$ , where the index  $a$  stands for different matter fields in the model.

Several new features of a supersymmetric field theories (as compared to a non-supersymmetric one) are evident from the above equation. First, note that the Yukawa couplings and the first

contribution to the Higgs potential arise from the same function  $W(\Phi)$  and therefore their parameters are related. The second term in the Higgs potential involves only the gauge coupling. Thus the number of parameters in a SUSY field theory are expected to be fewer than the usual gauge theories. Secondly, there is a new kind of Yukawa coupling involving the gaugino, matter fermion  $\psi$  and the superpartner of the matter field with the Yukawa coupling given by the gauge coupling constant  $g$ . These generic features have many important implications for the phenomenology of these models.

To see one immediate phenomenological implication of supersymmetry, consider a superpotential  $W(\Phi) = m\Phi^2$ . Using Eq.(8), it is very easy to see that the scalar field  $\phi$  and the fermion field  $\psi$  both have the same mass. This is again a generic prediction of the SUSY models, i.e. the particle and its superpartner have the same mass in the supersymmetric limit. This is clearly against observations for all known particles implying that in a realistic model, supersymmetry must be broken. As in the case of other global symmetries, supersymmetry breaking can be explicit or spontaneous. In the latter case, the analogs of Nambu-Goldstone theorem and the Higgs-Kibble mechanism imply that if supersymmetry is global and is spontaneously broken (i.e.,  $Q_{susy}|0\rangle \neq 0$ ), then there must exist a massless fermion in the particle spectrum (to be called the Goldstino) and it will obey low energy theorems — the analog of Adler's zeros for the pion. Since no particle with these properties is known to exist in nature, we have to assume that supersymmetry is either explicitly broken or more elegantly, supersymmetry is a local symmetry which is spontaneously broken. In the latter case, the analog of the Higgs-Kibble mechanism for local symmetries leads to the conclusion that the Goldstino becomes the longitudinal mode of the gauge particle corresponding to the local supersymmetry<sup>2</sup>. We will discuss this further in a subsequent section.

## 4 Some properties of exact N=1 supersymmetric theories

In this section, we review some of the new properties of supersymmetric field theories. We will only state the results without giving the detailed mathematical proofs and we refer the reader to the original literature for the proofs.

### IVa. Non-renormalization theorem:

It is well known that the only successful way to get useful information from ordinary local field theories is to work in the limit of weak couplings (typically,  $\lambda^2/4\pi \ll 1$ , where  $\lambda$  is an interaction strength in the theory) and use the perturbative solutions of the field equations. The singularity of the products of field operators then leads to divergences when one goes beyond the tree level approximations (i.e. to include Feynman diagrams that includes loops). This leads to a division of field theories into two classes: one class, known as renormalizable field theories is where the number of operators that emerge from loop calculations with divergences is finite to all orders in perturbation theory and the other class is known as nonrenormalizable theories and consists of theories where the number of operators that are accompanied by infinities proliferates without limit. In applications to physics applications (such as in particle physics) one chooses renormalizable field theories since in these theories, all infinities can be absorbed by redefining a finite number of observable parameters which then become the input parameters of the theory. They receive infinite renormalization in each loop order and if it is assumed that theory is an "effective" theory valid below some scale, the infinity is replaced by a dependence on the cutoff scale  $\Lambda$  below which the "effective" theory holds. Their values therefore cannot be chosen at the tree level. In such theories, however, there are other physical quantities that can be predicted in terms of these input parameters. This provides a way to experimentally test whether a given field

<sup>2</sup>Recall that Higgs-Kibble mechanism for ordinary continuous local symmetries refers to the phenomenon that if the symmetry is spontaneously broken, the associated massless Nambu-Goldstone boson becomes the longitudinal mode of the vector gauge boson (which in the beginning had only two states of polarization and was massless like the photon) to make it a massive gauge boson.



theory describes natural phenomena. The most illustrious example of this kind of theory is the quantum electrodynamics where the input parameters are the electron mass and charge and most recently the standard model of electroweak interactions which has more input parameters. In these models, the higher order corrections have actually been experimentally measured (e.g.  $g - 2$  of muons and electrons) proving them as fundamental theories of nature.

While the above argument implies that the nonrenormalizable theories may not have anything to do with reality, this need not necessarily be the case. For instance, it may be that a nonrenormalizable theory may describe phenomena observed in nature, a well known example being Einstein's theory of general relativity. In such cases, one may hope that they are "effective" field theories which are "low" energy manifestations of some deeper theory. One should not therefore discard a theory as uninteresting just because it is apparently nonrenormalizable.

As hinted earlier, supersymmetric theories by virtue of the extra symmetries in them have much "softer" divergence structure. A very important property of supersymmetric field theories is their ultraviolet behavior. One particular aspect of this softer divergence structure is the extremely important property that in the exact supersymmetric limit, the parameters of the superpotential  $W(\Phi)$  do not receive any (finite or infinite) corrections from Feynman diagrams involving the loops. In other words, if the value of a superpotential parameter is fixed at the classical level, it remains unchanged to all orders in perturbation theory. Furthermore, if  $W(\Phi)$  is a polynomial of order less than or equal to three, then there is only one divergent counter term in the theory which is absorbed by a redefinition of the kinetic energy term in the Lagrangian. Thus theory is renormalizable in the conventional sense but with only one class of infinities. Note of course that if  $W(\Phi)$  is at most a cubic polynomial in  $\Phi$ , then the corresponding field theory is a renormalizable field theory. This is known as the non-renormalization theorem [5]. This makes supersymmetric field theories far more appealing as theories of nature since the input parameters of the theory do not have to be defined at every loop order and everytime we encounter a threshold for new physics.

This observation was realized as the key to solving the Higgs mass problem of the standard model as follows: the radiative corrections to the Higgs mass in the standard model are quadratically divergent and admit the Planck scale as a natural cutoff if there is no new physics upto that level. Since the Higgs mass is directly proportional to the mass of the  $W$ -boson, the loop corrections would push the  $W$ -boson mass to the Planck scale destabilizing the standard model. On the other hand in the supersymmetric version of the standard model (to be called MSSM), in the limit of exact supersymmetry, there are no radiative corrections to any mass parameter and therefore to the Higgs boson mass which is a parameter of the superpotential. Thus if the world could be supersymmetric at all energy scales, the weak scale stability problem would be easily solved. However, since supersymmetry must be a broken symmetry, one has to ensure that the terms in the hamiltonian that break supersymmetry do not spoil the non-renormalization theorem in a way that infinities creep into the self mass correction to the Higgs boson. This is precisely what happens if effective supersymmetry breaking terms are "soft". We will discuss the soft breaking of supersymmetry in the next section.

As a final addendum to the discussion on the renormalization theorem, we note that if  $W(\Phi)$  is a polynomial involving fourth or higher order terms, then the corresponding field theory is nonrenormalizable. However, something very interesting happens in this case too. What happens is that the parameters of the superpotential (all of them) remain unaffected by the loop corrections. However, there are still infinite number of divergent counter terms which in the superfield language involve terms necessarily of the form  $\Phi^\dagger \Phi \Phi^n$ . In the ordinary field theory language, all counter terms involve space-time derivatives of the fields.

#### *IVb. Effective field theory of composite states in $N = 1$ super-Yang-Mills theories:*

The fundamental theory of strong interactions is now known to be described by an unbroken non-abelian gauge theory known as the Quantum Chromodynamics. It is an  $SU(3)$  gauge theory under which quarks and antiquarks transform as  $\mathbf{3}$  and  $\mathbf{3}^*$  dimensional representations. The

baryons (protons and neutrons) and mesons observed in nature are assumed to be color (i.e.  $SU(3)_c$ ) singlet bound states of this underlying theory, when the strong interactions become strong. In order to understand the dynamics of nuclear forces and the systems of baryons and mesons, one needs an effective low energy Lagrangian involving the mesons and the baryons. It has been argued by Weinberg and others that one way to arrive at such effective Lagrangians is to use the fact that the underlying QCD theory is invariant under the chiral symmetry i.e.  $SU(3)_L \times SU(3)_R$  and the spectrum of pseudoscalar mesons implies that this symmetry is spontaneously broken by the ground state of the theory. This assumption is very helpful in understanding the dynamics of mesons and baryons. While there are various plausible arguments that can be given in favor of the spontaneous breaking of chiral symmetry in QCD, it has never been satisfactorily established. Furthermore, the effective Lagrangians written down on the basis of symmetry arguments are at best approximate. The question then arises as to whether there exist any gauge theories that resemble QCD whose low energy Lagrangian can be written down exactly and if so can questions like whether spontaneous breaking of chiral symmetry actually occurs, be answered.

It has been recently shown[6] that for  $N=1$  supersymmetric Yang-Mills theories, an exact effective Lagrangian can be written down for the composite color singlet states. This raises the hope that approaching QCD from this angle may provide answers to some of the above questions and a better understanding of the dynamics of baryons and mesons.

It may be instructive to give an example of this class of models. Consider an  $N = 1$   $SU(N_c)$  gauge theory with left and right handed "quarks" (denoted by  $Q$  and  $Q^c$ ) transforming as  $N_c$  and  $N_c^*$  representations of the group. Consider  $N_f$  families of such representations with  $N_f < N_c$ . This model has the symmetry  $G \equiv SU(N_f)_L \times SU(N_f)_R \times U(1)_A \times U(1)_B \times U(1)_R$  under which the matter fields transform as:  $Q (N_f, 1, 1, 1, \frac{N_f - N_c}{N_f})$  and  $Q^c (1, N_f^*, 1, -1, \frac{N_f - N_c}{N_f})$ . Here the first four groups are familiar hadronic symmetry groups i.e. chiral  $SU(3)_L \times SU(3)_R$  (for three quark flavors); axial  $U(1)$  and baryon number symmetry whereas the last  $U(1)_R$  is typical of supersymmetry under which not only the fields but also the supersymmetry coordinate  $\theta$  transform nontrivially. It is the presence of this extra symmetry that helps to restrict the form of the composite field theory below the scale  $\Lambda$  of the nonabelian gauge theory defined by the mass below which the coupling becomes strong and composites form. The argument is that whatever the form of the composite field theory is, it must respect the above symmetry (in much the same fashion that one constructs the pion Lagrangian in the ordinary QCD). Furthermore since we are working in the supersymmetric limit, the form of the effective Lagrangian must be supersymmetric i.e. it must have a superpotential that is a holomorphic function of the composite fields which themselves must be  $SU(N_c)$  invariant. These restrictions i.e. invariance  $G$  and holomorphy is satisfied for the case of  $N_f < N_c$  by only one function of the only  $SU(N_c)$  invariant composite field  $Q^c Q$  i.e.

$$W_{eff} = C_{N_c, N_f} \left( \frac{\Lambda^{3N_c - N_f}}{\det Q^c Q} \right)^{\frac{1}{(N_c - N_f)}} \quad (9)$$

where  $C_{N_c, N_f}$  are constants which depend on the details of renormalization[7]. It is very interesting that the effective composite particle theory has a unique superpotential. There could of course be an arbitrary form for the Kahler potential (i.e. terms of the form  $\Phi^\dagger \Phi \Phi^n$ ). Since the ground state of the theory is determined by the superpotential of the theory, this form is extremely helpful in understanding the nature of the composite particle theory and has been used in the literature to build composite models of quarks and leptons[8].

When one applies the above considerations to the case where  $N_f = N_c$ , one finds that the resulting composite field theory breaks the chiral symmetry  $SU(N_f)_L \times SU(N_f)_R$  down to its vector sum group  $SU(N_f)_V$  exactly as in the case of QCD. Thus one of the key properties of known strong interactions i.e. spontaneous breakdown of chiral symmetries is explained by this specific supersymmetric model. It must be cautioned that we do not know whether these properties will survive to the case real QCD which does not have supersymmetry.

#### IVc. Vanishing cosmological constant

Observations of the various cosmological phenomena seem to indicate that the universe has either zero or an extremely tiny cosmological constant. (The recent type I supernovae observations seem to imply a cosmological constant,  $\sim (10^{-3} \text{ eV})^4$ .) On the other hand, spontaneously broken field theories which seem to be necessary to describe weak and electromagnetic interactions imply a large cosmological constant ( $\sim (246 \text{ GeV})^4$ , which is  $10^{56}$  orders larger than the observed value). Similarly, QCD phase transitions also seem to imply large cosmological constants. How does one reconcile these theories with observations. Here supersymmetry provide a way out although no realistic construction of such theories have been made.

The reason for this hope is to note that a cosmological constant in the field theories is the value of the potential in the ground state. On the other hand in supersymmetric field theories, the potential is gives by

$$V(\phi) = |W_a|^2 + \frac{1}{2} \mathcal{D}_\alpha \mathcal{D}_\alpha \quad (10)$$

Note that both the terms are positive definite and therefore the minimum of this potential will always correspond to  $V(\Phi) = 0$ . Thus supersymmetry always implies a vanishing cosmological constant almost trivially.

Of course, as we will discuss, in real life, supersymmetry is always broken; so the question then is how these properties get modified. As we will discuss, most of the power of the nonrenormalization theorem is maintained even in the presence of a certain class of supersymmetry breaking terms known as “soft” breaking terms. As far as the form of the low energy Lagrangian in the SUSY Yang-Mills theories goes, however, the situation is model dependent. If supersymmetry is broken spontaneously, then one can use the effective field theory approach. In other cases, the situation is unknown. Finally, the vanishing of cosmological constant can not be maintained in a natural manner once supersymmetry is broken.

## 5 Local supersymmetry and Gravity

Local symmetries seem to have played a key role in our understanding of the nature of the fundamental interactions e.g.  $SU(2)_L \times U(1)_Y$  for the case of electroweak interactions and  $SU(3)_c$  for the case of strong interactions etc. There also exists a formulation of gravity as a local symmetry of translation and Lorentz invariance. It is therefore natural to ask that if supersymmetry is to play a role in understanding the puzzles of the standard model (as for instance the Higgs mass alluded to before), could it be a local symmetry? It actually turns out that this question is more than matter mere curiosity, as was realized in the late seventies[9].

As is well known, corresponding to every local symmetry, there is a gauge field. In the case of supersymmetry, the corresponding gauge field must have spin  $\frac{3}{2}$ . One must therefore look for a complete supersymmetric theory involving the spin  $\frac{3}{2}$  field. A very exciting aspect of local supersymmetry is that supermultiplet containing the above spin  $\frac{3}{2}$  field (to be called gravitino henceforth) contains the graviton (which has spin 2) as its superpartner. Of course one could have thought that may be instead of a spin 2 particle, one could use a spin 1 boson to complete the gravitino supermultiplet and we would not then have gravity in the theory. To see why one must necessarily have gravity in the gauge multiplet of local supersymmetry, not that the supersymmetry algebra involves the translation operator in the right hand side. Therefore, if we gauge supersymmetry, we must at least gauge translation. As we just noted, the gauge field corresponding to local translation is the vierbein  $e_\mu^a$ . Thus local supersymmetry automatically leads to a supersymmetric theory that includes gravity. In fact elementary arguments can be used to infer that unlike the usual gauge symmetries, the coupling constant corresponding to local supersymmetry has mass dimension  $-1$  and leads to a natural identification with  $\sqrt{G_N} \equiv \kappa \equiv M_{Pl}^{-1}$ . To see this explicitly, let us first recall how the gauge covariant derivative is written down for an ordinary local symmetry, say an  $U(1)$  symmetry. Suppose there is a scalar field which has charge one under the local symmetry. Then under the local symmetry transformation, we have change in

the field  $\phi$  is given by

$$\delta\phi = i\alpha(x)\phi \quad (11)$$

where  $\alpha$  is the infinitesimal parameter of the gauge transformation. The gauge covariant derivative is then given by

$$D_\mu\phi(x) = \partial_\mu\phi - igA_\mu\phi \quad (12)$$

Here  $g$  is the gauge coupling constant. Note that since  $\phi$ ,  $A_\mu$  and  $\partial$  all have mass dimension 1, simple dimensional counting tells us that  $g$  is dimensionless.

Now let us apply the similar considerations to local supersymmetry. Under local supersymmetry transformation, the members of a supermultiplet  $(\phi, \psi, F)$ , transform among each other. We can write them as follows:

$$\begin{aligned} \delta\phi &= \epsilon\psi \\ \delta\psi &= \epsilon\sigma^\mu\partial_\mu\phi + \epsilon F \\ \delta F &= \partial_\mu\psi\sigma^\mu\bar{\epsilon} \end{aligned} \quad (13)$$

If we want to write the covariant derivative for the scalar field for the case of local supersymmetry, then following the  $U(1)$  case, we would write

$$D_\mu = \partial_\mu - i\kappa\psi\psi_\mu \quad (14)$$

where  $\psi_\mu$  is the gravitino field. Now in order to carry out dimensional counting note that all fermions have mass dimension  $\frac{3}{2}$  from which it follows that  $\kappa$ , the gauge coupling of local supersymmetry has dimension of inverse mass since both terms in the covariant derivative must have the same mass dimension. And as we just noted  $\kappa = \sqrt{G}$  where  $G$  is Newton's constant.

One can write down the generalization of the Lagrangian invariant under local supersymmetry[9]. An important point to note is that local supersymmetry brings in a new massless particle with spin  $\frac{3}{2}$ , the gravitino which couples to all kinds of matter by virtue of the fact that it is the superpartner of the graviton and graviton of course couples to all matter. However, since the gravitino coupling is very weak i.e. of order  $\sqrt{G}$ , its presence will not be felt very "strongly" by experiments. Nonetheless its presence gives rise to many new kinds of physical as well as astrophysical phenomena and is a very active area of investigation.

In the limit of exact supersymmetry, the gravitino like the graviton is massless. But it becomes massive once supersymmetry is broken (see later).

## 6 Breaking of supersymmetry

If supersymmetry was an exact symmetry of nature, we would have had equal mass superpartners for all known quarks, leptons and gauge bosons and they should have been discovered in the collider as well as many other experiments. However, superpartner of no known particle has been experimentally discovered. Therefore, like many other symmetries of physics, supersymmetry must be broken.

There are two ways to break a symmetry of nature: (i) by explicit nonsymmetric terms in the Lagrangian, as in the case of quark masses breaking global  $SU(2)$  or  $SU(3)$  symmetries among baryons and mesons; or (ii) by the vacuum state not being invariant under the symmetry. The latter mechanism is known as the Nambu-Goldstone (NG) way of breaking the symmetry and is always accompanied by the prediction of a massless state in the spectrum of states in the theory with the same quantum numbers as the broken generator of the symmetry. Since the generator of supersymmetry is a fermionic operator, the NG particle in this case must be a fermion. It is called a goldstino in the literature. Just as in the case of the NG bosons, the goldstino (denoted by the symbol  $\chi$ ) coupling to particles of a supermultiplet must obey fermionic shift invariance i.e. i.e.

$\chi \rightarrow \chi + \epsilon$  where  $\epsilon$  is a constant grassman number. If the scale of supersymmetry breaking is given by  $\Lambda$ , then a typical coupling of the goldstino will be of the form  $\frac{\partial_\mu \chi}{\Lambda^2}$ .

Again since no massless fermion has been discovered to date with properties identifiable with the gravitino, it is believed that there must be some mechanism that makes the gravitino massive. One simple way out is to invoke the analog of the Higgs mechanism for local SUSY which can turn the Goldstino into the longitudinal mode of the massive spin 3/2 gravitino. This way both the massless gravitino and the massless goldstino disappear from the spectrum of the theory. We will assume this to be a feature of the supersymmetric theories that we will consider.

Now let us note a very important feature that is analogous to the case of Higgs mechanism of normal local symmetries. In the case of the usual local symmetries, say, the U(1) symmetry discussed above, if there is spontaneous symmetry breaking of U(1) local symmetry, then the mass of the gauge field becomes  $g < \phi > \equiv gv$  where  $v$  is the scale of the symmetry breaking. In exact analogy, the mass of the gravitino will be  $\kappa \Lambda^2$  since  $\kappa$  is the analog of the gauge coupling  $g$  and  $\Lambda^2$  is the analog of  $< \phi >$ . Thus we see that if one could measure the mass of the gravitino, one would be able to determine the scale of the supersymmetry breaking.

The practical way to implement spontaneous breaking of supersymmetry is to follow the analogy with the spontaneous breaking of ordinary bosonic symmetries. Suppose one wants to break a U(1) symmetry spontaneously. The way to do is to consider a bosonic field  $\phi$  (or a composite bilinear of fermions) which carries nonzero charge under the U(1) symmetry and arrange the theory such that  $< \phi > \neq 0$ . In the case of supersymmetry, we have to look for bosonic fields that carry the supersymmetry charge (we need bosonic fields since we do not want to break Lorentz invariance) and give them non zero vevs. Two examples of bosonic fields with nonzero SUSY charge are the  $F$  field component of a chiral superfield and the D-component of a gauge supermultiplet. If one chooses a superpotential such that either of these fields can acquire vev, then supersymmetry will be spontaneously broken. Alternatively, if one adds a linear D-term to the theory, that will lead to nonvanishing value for the D-term in the ground state and will lead to breakdown of supersymmetry.

A simple example of the F-type supersymmetry breaking is given by the choice of a superpotential of the form:

$$W = \mu^2 z \quad (15)$$

Note that since the F-term is obtained by taking a derivative  $\partial W / \partial z$ , we see that  $F_z = \mu^2 \neq 0$ . Thus supersymmetry is broken. In this case the Goldstino field is the fermionic component of the superfield  $z$  i.e.  $\psi_z$  which is easily seen to be mass less.

Let us now focus on the other way of breaking supersymmetry i.e. by adding explicit supersymmetry breaking terms to the Lagrangian. If we want to maintain the good divergence structure of the theory subsequent to the addition of the SUSY breaking terms, new constraints must be satisfied by those terms - we will call the allowed terms of this restricted variety the soft SUSY breaking terms. Let us enumerate them below:

1.  $m_a^2 \phi_a^\dagger \phi_a$ , where  $\phi$  is the bosonic component of the chiral superfield  $\Phi_a$ ;
2.  $m \int d^2 \theta d^2 \bar{\theta} (A W^{(3)}(\Phi) + B W^{(2)}(\Phi))$ , where  $W^{(3)}(\Phi)$  and  $W^{(2)}(\Phi)$  are the second and third order polynomials in the superpotential.
3.  $\frac{1}{2} m_\lambda \lambda^T C^{-1} \lambda$ , where  $\lambda$  is the gaugino field.

It can be shown that the soft breaking terms only introduce finite loop corrections to the parameters of the superpotential. Since all the soft breaking terms require couplings with positive mass dimension, the loop corrections to the Higgs mass will depend on this mass and we must keep these masses less than a TeV so that the weak scale remains stabilized. This has the interesting implication that superpartners of the known particles must have masses in the range of 100 GeV to a TeV and are accessible to the ongoing and proposed collider experiments such as the Tevatron at Fermilab and LHC at CERN. For a recent survey of the experimental situation, see Ref. [10].

## 7 Origin of supersymmetry breaking

An extremely important question in the application of supersymmetry to particle physics focusses on the origin of the soft breaking terms. Their detailed pattern depends sensitively on the particular way that supersymmetry is broken. Thus by experimentally unravelling the pattern of the superpartner masses (i.e. squarks, sleptons, gauginos etc), one can throw light on the nature and the origin of supersymmetry breaking.

The usual strategy employed in implementing supersymmetry breaking is to assume that SUSY is broken in a hidden sector that does not involve any of the matter or forces of the standard model (which we call the visible sector) and this SUSY breaking is transmitted to the visible sector via some intermediary, to be called the messenger sector.

There are generally two ways to set up the hidden sector- a less ambitious one where one writes an effective Lagrangian (or superspotential) in terms of a certain set of hidden sector fields that lead to supersymmetry breaking in the ground state and another more ambitious one where the SUSY breaking arises from the dynamics of the hidden sector interactions. In implementing the second method, one uses the fact that in nonabelian  $N=1$  gauge models with matter, the composite field theory can be written down exactly in many cases, as discussed above and if the resulting field theory has a vacuum that breaks supersymmetry, that is taken as a prototype hidden sector.

For our purpose we will use the simpler schemes of the first kind. As far as the messenger sector goes there are three possibilities as already referred to earlier: (i) gravity mediated [11]; (ii) gauge mediated [12] and (iii) anomalous  $U(1)$  mediated [13]. Below we give examples of each class.

This sector is made somewhat more technical than the rest of the article on purpose so that any one "wishing to get his/her feet wet" in the supersymmetry game can actually proceed to that endeavour right after reading this article.

### VIIa. Gravity mediated SUSY breaking

The scenario that uses gravity to transmit the supersymmetry breaking is one of the earliest hidden sector scenarios for SUSY breaking and forms much of the basis for the discussion in current supersymmetry phenomenology. In order to discuss these models one needs to know the supergravity couplings to matter. An essential feature of supergravity coupling is the generalized kinetic energy term in gravity coupled theories called the Kahler potential,  $K$ . We will use a function  $G$  of the Kahler potential. It is a hermitean operator and a function of the matter fields in the theory and their complex conjugates. The effect of supergravity coupling in the matter and the gauge sector of the theory is given in terms of  $G$  and its derivatives as follows:

$$L(z) = G_{zz^*} |\partial_\mu z|^2 + e^{-G} [G_z G_{z^*} G_{zz^*}^{-1} + 3] \quad (16)$$

where  $z$  is the bosonic component of a typical chiral field (e.g. we would have  $z \equiv \bar{q}, \bar{l}$  etc) and  $G = 3 \ln(-\frac{K}{3}) - \ln|W(z)|^2$ . A superscript implies derivative with respect to that field. The

simplest choice for the Kahler potential  $K$  is  $K = -3e^{-\frac{|z|^2}{3M_{Pl}^2}}$  that normalizes the kinetic energy term properly. Using this, one can write the effective potential for supergravity coupled theories to be:

$$V(z, z^*) = e^{\frac{|z|^2}{M_{Pl}^2}} \left[ |W_z + \frac{z^*}{M_{Pl}^2} W|^2 - \frac{3}{M_{Pl}^2} |W|^2 \right] + D - terms \quad (17)$$

The gravitino mass is given in terms of the Kahler potential as :

$$m_{3/2} = M_{Pl} e^{-G/2} \quad (18)$$

A popular scenario suggested by Polonyi is based upon the following hidden sector consisting of a gauge singlet field, denoted by  $z$  and the superpotential  $W_H$  given by:

$$W_H = \mu^2(z + \beta) \quad (19)$$

where  $\mu$  and  $\beta$  are mass parameters to be fixed by various physical considerations. It is clear that this superpotential leads to an F-term that is always non-vanishing and therefore breaks supersymmetry. Requiring the cosmological constant to vanish fixes  $\beta = (2 - \sqrt{3})M_{Pl}$ . Given this potential and the choice of the Kahler potential as discussed earlier, supergravity calculus predicts a universal soft breaking parameters  $m$  given by  $m_0 \sim \mu^2/M_{Pl}$ . Requiring  $m_0$  to be in the TeV range implies that  $\mu \sim 10^{11}$  GeV. The complete potential to zeroth order in  $M_{Pl}^{-1}$  in this model is given by:

$$V(\phi_a) = [\Sigma_a |\frac{\partial W}{\partial \phi_a}|^2 + V_D] \quad (20)$$

$$+ [m_0^2 \Sigma_a \phi_a^* \phi_a + (AW^{(3)} + BW^{(2)} + h.c.)]$$

where  $W^{(3,2)}$  denote the dimension three and two terms in the superpotential respectively. The values of the parameters  $A$  and  $B$  at  $M_{Pl}$  are related to each other in this example as  $B = A - 1$ . The gaugino masses in these models arise out of a separate term in the Lagrangian depending on a new function  $f(z)$  of the hidden sector singlet fields,  $z$ :

$$\int d^4x d^2\theta f(z) W_\lambda^\alpha W_{\lambda,\alpha} \quad (21)$$

If we choose  $f(z) = \frac{z}{M_{Pl}}$ , then gaugino masses come out to be order  $m_{3/2} \sim \frac{\mu^2}{M_{Pl}}$  which is also of order  $m_0$ , i.e. the electroweak scale. Furthermore, in order to avoid undesirable color and electric charge breaking by the SUSY models, one must require that  $m_0^2 \geq 0$ .

It is important to point out that the superHiggs mechanism operates at the Planck scale. Therefore all parameters derived at the tree level of this model need to be extrapolated to the electroweak scale. So after the soft-breaking Lagrangian is extrapolated to the weak scale, it will look like:

$$\mathcal{L}^{SB} = m_a^2 \phi_a^* \phi_a + m \Sigma_{i,j,k} A_{ijk} \phi_i \phi_j \phi_k + \Sigma_{i,j} B_{ij} \phi_i \phi_j \quad (22)$$

We will see later that this extrapolation arising from the radiative corrections to the theory eventually leads to an understanding of the origin of the weak scale.

#### VIIb. Gauge mediated SUSY breaking[12]

This mechanism for the SUSY breaking has recently been quite popular in the literature and involves different hidden as well as messenger sectors. In particular, it proposes to use the known gauge forces as the messengers of supersymmetry breaking. As an example, consider a unified hidden messenger sector toy model of the following kind, consisting of the fields  $\Phi_{1,2}$  and  $\bar{\Phi}_{1,2}$  which have the standard model gauge quantum numbers and a singlet field  $S$  and with the following superpotential:

$$W = \lambda S(M_0^2 - \bar{\Phi}_1 \Phi_1) + M_1(\bar{\Phi}_1 \Phi_2 + \Phi_1 \bar{\Phi}_2) + M_2 \bar{\Phi}_1 \Phi_1 \quad (23)$$

The F-terms of this model are given by:

$$F_S = \lambda(M_0^2 - \bar{\Phi}_1 \Phi_1) \quad (24)$$

$$F_{\Phi_2} = M_1 \bar{\Phi}_1; \quad F_{\bar{\Phi}_2} = M_1 \Phi_1$$

$$F_{\Phi_1} = M_2 \bar{\Phi}_1 + M_1 \bar{\Phi}_2 - \lambda S \bar{\Phi}_1$$

It is easy to see from the above equation that for  $M_1 \gg M_0, M_2$ , the minimum of the potential corresponds to all  $\Phi$ 's having zero vev and  $F_S = \lambda M_0^2$ , thus breaking supersymmetry. The same superpotential responsible for SUSY breaking also transmits the SUSY breaking information to the visible sector. While the spirit of this model is similar to the original papers on the subject this unified construction is different and has its characteristic predictions.

The SUSY breaking is transmitted to the visible sector via one and two loop diagrams. The gaugino masses arise from the one loop diagram where a gaugino decomposes into the SUSY partners  $\phi_1$  and  $\bar{\phi}$  and the loop is completed as  $\phi_1$  and  $\bar{\phi}_1$  mix thru  $F_S$  susy breaking term and the fermionic partners mix via the mass term  $M_2$ . The squark and slepton masses arise from the two loop diagram where the squark-squark gauge boson -gauge boson coupling begins the first loop and one of the gauge bosons couples to the two  $\phi_1$ 's and another to the two  $\bar{\phi}_1$ 's which in turn mix via the F-terms for S to complete the two loop diagram. This is only one typical diagram and there are many more which contribute in the same order. It is then easy to see that their magnitudes are given by:

$$m_\lambda \simeq \frac{\alpha}{4\pi} \frac{\langle F_S \rangle}{M_2} \quad (25)$$

$$m_{\tilde{q}}^2 \simeq \left( \frac{\alpha}{4\pi} \right)^2 \left( \frac{\langle F_S \rangle}{M_2} \right)^2$$

The first point to notice is that the gaugino and squark masses are roughly of the same order and requiring the squark masses to be around 100 GeV, we get for  $F_S/M_2 \simeq 100$  TeV. Of course,  $\langle F_S \rangle$  and  $M_2$  need not be of same order in which case the numerics will be different.

A distinguishing feature of this approach is that due to low scale for SUSY breaking, the gravitino mass is always in the milli-eV to kilo-eV range and is therefore is always the LSP. Thus these models cannot provide a candidate for the cold dark matter of the universe, which was always considered an added attraction of supersymmetric models. The attractive property of these models is that they lead naturally to near degeneracy of the squark and sleptons thus alleviating the FCNC problem of the MSSM and have therefore been the focus of intense scrutiny during the past year.

These class of models however suffer from the fact that the messenger sector is too adhoc and practically an arbitrary number of models can be constructed just by varying the messenger sector. Nevertheless it is a very interesting class of models and should be tested experimentally.

#### VIIc. Anomalous $U(1)$ mediated supersymmetry breaking

These class of models owe their origin to the string models, which after compactification can often leave anomalous  $U(1)$  gauge groups[14]. Since the original string model is anomaly free, the anomaly cancellation must take place via the Green-Schwarz mechanism as follows. Consider a  $U(1)$  gauge theory with a single chiral fermion that carries a  $U(1)$  quantum number. This theory has an anomaly. Therefore, under a gauge transformation, the low energy Lagrangian is not invariant and changes as:

$$L \rightarrow L + \frac{\alpha}{4\pi} F \tilde{F} \quad (26)$$

where  $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$  and  $\tilde{F}$  is the dual of  $F_{\mu\nu}$ . The last term is the anomaly term. To restore gauge invariance, we can add to the Lagrangian the Green-Schwarz term and rewrite the effective Lagrangian as

$$L' = L + \frac{a}{M} F \tilde{F} \quad (27)$$

where under the gauge transformation  $a \rightarrow a - M\alpha/4\pi$ . In order to obtain the supersymmetric version of the Green-Schwarz term, we have to add a dilaton term to the axion  $a$  to make a complex chiral superfield. Let us denote the dilaton field by  $\phi$  and the complex chiral field containing it as  $S = \phi + ia$ . The gauge invariant action containing the  $S$  and the gauge supersfield  $V$  has terms of the following form:

$$A = \int d^4\theta \ln(S + S^\dagger - V) + \int d^2\theta SW^\alpha W_\alpha + \text{matter field parts} \quad (28)$$

It is clear that in order to get a gauge field Lagrangian out of this, the dialton  $S$  must have a vev with the identification that  $\langle S \rangle = g^{-2}$  and it is a fundamental unanswered question in



superstring theory as to how this vev arises. If we assume that this vev has been generated, then, one can see that the first term in the Lagrangian when expanded around the dilaton vev, leads to a term  $\frac{1}{\langle 2S \rangle} \int d^4\theta V$ , which is nothing but a linear Fayet-Illiopoulos D-term. Combining this with other matter field terms with non-zero U(1) charge, one can then write the D-term of the Lagrangian. As an example that can lead to realistic model building, we take two fields with equal and opposite U(1) charges  $\pm 1$  in addition to the squark and slepton fields. The D-term can then be written as:

$$V_D = \frac{g^2}{2} (n_q^2 |\tilde{Q}|^2 + n_L^2 |\tilde{L}|^2 + |\phi_+|^2 - |\phi_-|^2 + \zeta)^2 \quad (29)$$

This term when minimized does not break supersymmetry. However, if we add the superpotential a term of the form  $W_\phi = m\phi_+\phi_-$ , then there is another term in low energy effective potential that leads to the combined potential as:

$$V = V_D + m^2 (|\phi_+|^2 + |\phi_-|^2) \quad (30)$$

The minimum of this potential corresponds to:

$$\langle \phi_+ \rangle = 0; \langle \phi_- \rangle = (\zeta - \frac{m^2}{g^2})^{1/2} \simeq \epsilon M_{Pl} : F_{\phi_+} = m M_{Pl} \epsilon \quad (31)$$

where we have assumed that  $\zeta = \epsilon^2 M_{Pl}^2$ . This then leads to nonzero squark masses  $m_Q^2 \simeq n_Q^2 m^2$ . Thus supersymmetry is broken and superpartners pick up mass. In the simplest model it turns out that the gaugino masses may be too low and one must seek ways around this. However, the A and B-terms are also likely to be small in this model and that may provide certain advantages. On the whole, this approach has great potential for model building and has not been thoroughly exploited[13]- for instance, it can be used to solve the FCNC problems, SUSY CP problem, to study the fermion mass hierarchies etc. It is beyond the scope of this review to enter into those areas.

## 8 Supersymmetric standard model

We will now apply the discussions of the previous sections to construct the supersymmetric extension of the standard model so that the goal of stabilizing the Higgs mass is indeed realized in practice. Before we do that let us briefly summarize the salient features of the successful standard model.

The standard model of electroweak and strong interactions based on the local symmetry group  $SU(3)_c \times SU(2)_L \times U(1)_Y$  is one of the major milestones of twentieth century theoretical physics. It not only provides a complete description of all known phenomena in particle physics in terms of a renormalizable field theory but it reaffirms the crucial rôle played by the local symmetry in the domain of weak and strong interactions and extends its success from Quantum Electrodynamics to Quantum Flavor Dynamics for weak interactions and Quantum Chromodynamics for strong interactions. The field content of the standard model is given in table I for the sake of completeness of this review.

**Table I.** The particle content of the standard model.

Field	gauge transformation
Quarks $Q_L$	$(3, 2, \frac{1}{6})$
Righthanded up quarks $u_R$	$(3, 1, \frac{2}{3})$
Righthanded down quarks $d_R$	$(3, 1, -\frac{2}{3})$
Lefthanded Leptons $L$	$(1, 2, -\frac{1}{2})$
Righthanded leptons $e_R$	$(1, 1, -1)$
Higgs Boson $H$	$(1, 2, +\frac{1}{2})$
Color Gauge Fields $G_a$	$(8, 1, 0)$
Weak Gauge Fields $W^\pm, Z, \gamma$	$(1, 3 + 1, 0)$

After spontaneous symmetry breaking, the  $W^\pm$  and the  $Z$  field acquire mass as do the fermions. The form of the neutral current interactions as well as the quantum corrections to this theory have been confirmed. The only particle of the standard model yet to be discovered is the Higgs boson and supersymmetry provides one class of extensions of the model where the Higgs boson has definite and testable properties.

There are however many unsolved problems in the standard model that for a long time had made many people suspect that there is an arena (or many arenas) of rich new physics beyond the standard model that would not only solve the puzzles of the standard model but also provide deeper insight into the prospects for unification of matter and forces.

The two puzzles that supersymmetry addresses successfully are the so called Higgs mass problem discussed already and the problem of understanding the gauge symmetry breaking. To discuss this, we start with a presentation of the particle contents of the model in detail as given in the table below.

**Table II caption.** The particle content of the supersymmetric standard model. For matter and Higgs fields, we have shown the left-chiral fields only. The right-chiral fields will have a conjugate representation under the gauge group.

Superfield	Particles	Superpartners	gauge transformation
Quarks $Q$	$(u, d)$	$(\tilde{u}, \tilde{d})$	$(3, 2, \frac{1}{3})$
Antiquarks	$u^c$	$\tilde{u}^c$	$(3^*, 1, -\frac{4}{3})$
Antiquarks	$d^c$	$\tilde{d}^c$	$(3^*, 1, \frac{2}{3})$
Leptons $L$	$(\nu, e)$	$(\tilde{\nu}, \tilde{e})$	$(1, 2, -1)$
Antileptons	$e^c$	$\tilde{e}^c$	$(1, 1, 2)$
Higgs Boson $H_u$	$(H_u^+, H_u^0)$	$(\tilde{H}_u^+, \tilde{H}_u^0)$	$(1, 2, +1)$
Higgs Boson $H_d$	$(H_d^0, H_d^-)$	$(\tilde{H}_d^0, \tilde{H}_d^-)$	$(1, 2, -1)$
Color Gauge Fields	$G_a$	$\tilde{G}_a$	$(8, 1, 0)$
Weak Gauge Fields	$W^\pm, Z$	$\tilde{W}^\pm, \tilde{Z}$	
Photon	$\gamma$	$\tilde{\gamma}$	

First note that an important difference between the standard model and its supersymmetric version apart from the presence of the superpartners is the presence of a second Higgs doublet. This is required both to give masses to quarks and leptons as well as to make the model anomaly free. The gauge interaction part of the model is easily written down following the rules laid out in [2]. In the weak eigenstate basis, weak interaction Lagrangian for the quarks and leptons is exactly the same as in the standard model. As far as the weak interactions of the squarks and the sleptons is concerned, the generation mixing angles are very different from those in the corresponding fermion sector due to supersymmetry breaking. This has the phenomenological implication that the gaugino-fermion-sfermion interaction changes generation leading to potentially large flavor changing neutral current effects such as  $K^0-\bar{K}^0$  mixing,  $\mu \rightarrow e\gamma$  decay etc unless the sfermion masses of different generations are chosen to be very close in mass.

Let us now proceed to a discussion of the superpotential of the model. It consists of two parts:

$$W = W_1 + W_2, \quad (32)$$

where

$$\begin{aligned} W_1 &= h_t^{ij} e_i^c L_j H_d + h_d^{ij} Q_i d_j^c H_d + h_u^{ij} Q_i u_j^c H_u + \mu H_u H_d \\ W_2 &= \lambda_{ijk} L_i L_j e_k^c + \lambda'_{ijk} L_i Q_j d_k^c + \lambda''_{ijk} u_i^c d_j^c d_k^c, \end{aligned} \quad (33)$$

$i, j, k$  being generation indices. We first note that the terms in  $W_1$  conserve baryon and lepton number whereas those in  $W_2$  do not. The latter are known as the  $R$ -parity breaking terms where  $R$ -parity is defined as

$$R = (-1)^{3(B-L)+2S}, \quad (34)$$

where  $S$  is the spin of the particle. It is interesting to note that the  $R$ -parity symmetry defined above assigns even  $R$ -parity to known particles of the standard model and odd  $R$ -parity to their superpartners. This has the important experimental implication that for theories that conserve  $R$ -parity, the super-partners of the particles of the standard model must always be produced in pairs and the lightest superpartner must be a stable particle. This is generally called the LSP. If the LSP turns out to be neutral, it can be thought of as the dark matter particle of the universe. In the presence of  $R$ -parity breaking interactions however, the LSP decays and the model has no cold dark matter candidate.

We now embed this model into the minimal  $N = 1$  supergravity model with a Polonyi type hidden sector. As a result, we get the mass splitting for the squarks and sleptons from the quarks and the leptons. We also get trilinear scalar interactions among the sfermions as follows:

$$\begin{aligned} \mathcal{L}^{SB} = & m_{3/2}[A_{e,ab}\tilde{e}_a^c\tilde{L}_bH_d + A_{d,ab}\tilde{Q}_aH_d\tilde{d}_b^c + A_{u,ab}\tilde{Q}_aH_u\tilde{u}_b^c] \\ & + B\mu m_{3/2}H_uH_d + \sum_{i=\text{scalars}} \mu_i^2\phi_i^\dagger\phi_i + \sum_a \frac{1}{2}M_a\lambda^TC^{-1}\lambda_a \end{aligned} \quad (35)$$

There will also be the corresponding terms involving the  $R$ -parity breaking interactions, which we omit here for simplicity.

The solution of the Higgs mass problem in the MSSM comes about as follows: if its tree level value is chosen to be of the order of the electroweak scale, any radiative correction to it will only induce terms of order  $\sim \frac{f^2}{16\pi^2}M_{SUSY}^2$ . By choosing the supersymmetry breaking scale in the TeV range as we did above, we can guarantee that to all orders in perturbation theory the Higgs mass remains stable and near the weak scale. Crucial to this result is the important property of nonrenormalization theorem[5] discussed earlier.

The way supersymmetry solves the problem of weak symmetry breaking is as follows. In order to have a realistic model, first a mechanism for supersymmetry breaking following one of the three ways already described is introduced. As is clear from the discussion of supersymmetry breaking, the hidden sector supersymmetry breaking manifests itself as a positive (mass)<sup>2</sup>'s for all scalar fields at the scale of SUSY breaking. (Some or all of them may also be equal depending on other details.) In order to study the theory at the weak scale, one must extrapolate all these parameters using the renormalization group equations. The degree of extrapolation will of course depend on the strength of the gauge and the Yukawa couplings of the various fields. In particular, the  $m_{H_u}^2$  will have a strong extrapolation proportional to  $\frac{h_t^2}{16\pi^2}$  since  $H_u$  couples to the top quark. Since  $h_t \simeq 1$ , this can make  $m_{H_u}^2(M_Z) < 0$ , leading to spontaneous breakdown of the electroweak symmetry. An approximate solution of the renormalization group equations gives

$$m_{H_u}^2(M_Z) = m_{H_u}^2(\Lambda_{SUSY}) - \frac{3h_t^2m_t^2}{16\pi^2} \ln \frac{\Lambda_{SUSY}^2}{M_Z^2} \quad (36)$$

This is a very attractive feature of supersymmetric theories.

Once the symmetry breaking has been implemented, constraints of supersymmetry provide one prediction that can distinguish the MSSM from the standard model- i.e. the mass of the lightest Higgs boson. It can be shown that the lightest higgs boson mass-square is going to be of order  $\sim g^2v_w^2$ . In fact denoting the vev's of the two Higgs doublets as  $\langle H_u^0 \rangle = v_u$  and  $\langle H_d^0 \rangle = v_d$ , one can write:

$$m_h^2 \simeq \frac{g^2 + g'^2}{4}(v_d^2 - v_u^2) \quad (37)$$

Defining  $v_u/v_d = \tan\beta$ , we can rewrite the above light Higgs mass formula as  $m_h^2 = M_Z^2 \cos 2\beta$  which implies that the tree level mass of the lightest Higgs boson is less than the  $Z$  mass. Once radiative corrections are taken into account,  $m_h$  increases above the  $M_Z$ . However, it is now well established that in a large class of supersymmetric models (which do not differ too much from the MSSM), the Higgs mass is less than 130 GeV or so.

## 9 Why go beyond the MSSM ?

Even though the MSSM solves two outstanding problems of the standard model, i.e. the stabilization of the Higgs mass and the breaking of the electroweak symmetry, it brings in a lot of undesirable consequences. They are:

(a) Presence of arbitrary baryon and lepton number violating couplings i.e. the  $\lambda$ ,  $\lambda'$  and  $\lambda''$  couplings described above. In fact a combination of  $\lambda'$  and  $\lambda''$  couplings lead to proton decay. Present lower limits on the proton lifetime then imply that  $\lambda'\lambda'' \leq 10^{-25}$  for squark masses of order of a TeV. Recall that a very attractive feature of the standard model is the automatic conservation of baryon and lepton number. In this sense therefore MSSM takes us one step backward from the standard model. The presence of R-parity breaking terms[15] also makes it impossible to use the LSP as the Cold Dark Matter of the universe since it is not stable and will therefore decay away in the very early moments of the universe. In various grand unified theories, keeping the R-parity violating terms under control provides a major constraint on model building.

(b) The different mixing matrices in the quark and squark sector leads to arbitrary amount of flavor violation manifesting in such phenomena as  $K_L - K_S$  mass difference etc. Using present experimental information and the fact that the standard model more or less accounts for the observed magnitude of these processes implies that there must be strong constraints on the mass splittings among squarks. Detailed calculations indicate that one must have  $\Delta m_{\tilde{q}}^2/m_{\tilde{q}}^2 \leq 10^{-3}$  or so. Again recall that this undoes another nice feature of the standard model where understanding the suppression of flavor violation was much simpler.

(c) The presence of new couplings involving the super partners allows for the existence of extra CP phases. In particular the presence of the phase in the gluino mass leads to a large electric dipole moment of the neutron unless this phase is assumed to be suppressed by two to three orders of magnitude. This is generally referred to in the literature as the SUSY CP problem. In addition, there is of course the famous strong CP problem which neither the standard model nor the MSSM provide a solution to.

(d) Finally there is a new naturalness problem that arises in the MSSM having to do with the  $H_u H_d$  term in the superpotential (see Eq. (33))- the so called  $\mu$  term. The point is that adequate electroweak symmetry breaking requires that this be of order of a few hundred GeV at most. But since this parameter is allowed in the supersymmetry limit, there is no reason for it to be smaller than say Planck mass. Although there are no divergent corrections to this parameter and in that sense it is different from the Higgs mass conundrum of the standard model, nonetheless the fact that we have to choose it to be small when a priori it could have been huge is not very satisfactory. It is hoped that in future theories that go beyond the MSSM this problem, known as the  $\mu$  problem will be addressed.

In order to cure these problems, one must seek new physics beyond the MSSM. Below, we discuss the example of the supersymmetric left-right model, which leads to automatic B and L conservation as well as solves the SUSY CP problem. These models also provide a solution to the strong CP problem without the need for an axion under certain circumstances.

## 10 Supersymmetric Left-Right model

The gauge group for this model[16] is  $SU(2)_L \times SU(2)_R \times U(1)_{B-L} \times SU(3)_c$ . The chiral superfields denoting left-handed and right-handed quark superfields are denoted by  $Q \equiv (u, d)$  and  $Q^c \equiv (u^c, d^c)$  respectively and similarly the lepton superfields are given by  $L \equiv (\nu, e)$  and  $L^c \equiv (\nu^c, e^c)$ [17]. The  $Q$  and  $L$  transform as left-handed doublets with the obvious values for the  $B-L$  and the  $Q^c$  and  $L^c$  transform as the right-handed doublets with opposite  $B-L$  values. The symmetry breaking is achieved by the following set of Higgs superfields:  $\phi_a(2, 2, 0, 1)$  ( $a = 1, 2$ );  $\Delta(3, 1, +2, 1)$ ;  $\bar{\Delta}(3, 1, -2, 1)$ ;  $\Delta^c(1, 3, -2, 1)$  and  $\bar{\Delta}^c(1, 3, +2, 1)$ . Unlike in the MSSM, the allowed terms in the superpotential are very limited in this case:

$$\begin{aligned}
 W = & h_a Q \phi_a Q^c + h'_a L \phi_a L^c + f(LL\Delta + L^c L^c \Delta^c) \\
 & + \mu_{ab} Tr(\phi_a \phi_b) + M(\Delta \bar{\Delta} + \Delta^c \bar{\Delta}^c)
 \end{aligned}
 \tag{38}$$

It is clear from the above equation that this theory has no baryon or lepton number violating terms and it allows for a dark matter particle.

The next question is how one breaks the  $SU(2)_R$  symmetry so that the successes of the standard model including the observed predominant V-A structure of weak interactions at low energies is reproduced. Another question of naturalness that also arises simultaneously is that since the charged fermions and the neutrinos are treated completely symmetrically (quark-lepton symmetry) in this model, how does one understand the smallness of the neutrino masses compared to the other fermion masses.

It turns out that both the above problems of the LR model have a common solution. The process of spontaneous breaking of the  $SU(2)_R$  symmetry that suppresses the V+A currents at low energies also solves the problem of ultralight neutrino masses. To see this let us specify the Higgs sector of the model: we choose  $SU(2)$  triplets with  $B - L = 2$  in left-right symmetric pairs ( $\Delta_L \oplus \Delta_R$ ) and bidoublets ( $\phi$ ) with  $B - L = 0$  which are left-right symmetric generalizations of the standard model Higgs doublet. The various components of the above Higgs fields is given as:

$$\Delta = \begin{pmatrix} \Delta^+/\sqrt{2} & \Delta^{++} \\ \Delta^0 & -\Delta^+/\sqrt{2} \end{pmatrix}; \quad \phi = \begin{pmatrix} \phi_1^0 & \phi_2^+ \\ \phi_1^- & \phi_2^0 \end{pmatrix} \quad (39)$$

The  $SU(2)_R \times U(1)_{B-L}$  symmetry is broken by giving a large vev to the  $\Delta^{c,0}$  field (i.e.  $\langle \Delta^{c,0} \rangle = v_R$ ). If  $v_R \gg m_{W_L}$ , then the V+A current effects at low energies are suppressed compared to the observed V-A current effects observed in beta decay. At the same time the  $L^c L^c \Delta^c$  coupling in the superpotential for the left-right model leads to a mass for the righthanded neutrino which is given by  $m_N = f v_R$ . The lefthanded neutrinos at this stage are massless. In the two component notation for the neutrino, this leads to the following mass matrix for the  $\nu, N$  (where we have denoted the left handed neutrino by  $\nu$  and the right handed component by  $N$ ).

$$M = \begin{pmatrix} 0 & h\kappa \\ h\kappa & f v_R \end{pmatrix} \quad (40)$$

By diagonalizing this  $2 \times 2$  matrix, we get the light neutrino eigen value to be  $m_\nu \simeq \frac{(h\kappa)^2}{f v_R}$ . Note that typical charged fermion masses are given by  $h'\kappa$  etc. So since  $v_R \gg \kappa, \kappa'$ , the light neutrino mass is automatically suppressed. This way of suppressing the neutrino masses is called the seesaw mechanism [18]. Thus in one stroke, one explains the smallness of the neutrino mass as well as the suppression of the V+A currents.

Another attractive feature of the LR models is that constraints of parity symmetry (under which  $Q_L \rightarrow Q_R, \phi \rightarrow \phi^\dagger$  etc), require the Yukawa couplings  $h_a, h'_a$  to be hermitean. As a result, if the vacuum of the theory is such that the  $\langle \phi \rangle$  vevs are real, then the quark mass matrices are hermitean. This means that the strong CP parameter  $\bar{\Theta}$  which is equal to  $Arg(Det M_q)$  vanishes at the tree level. This point was emphasized in the late 70's as a possible way to solve the strong CP problem without introducing the axion. The only problem that was noted with this suggestion at that time was that without imposing extra symmetries, it was not possible to have the  $\langle \phi \rangle$  to be real. This problem gets cured automatically once the left-right model is made supersymmetric.

Furthermore parity symmetry also makes the gluino mass,  $\mu$  and  $B\mu$  terms all real unlike the situation in MSSM. This solves the so called supersymmetric CP problem of the MSSM.

This model however does not throw any light on either the  $\mu$  problem or the FCNC problem. They will perhaps have a solution of supersymmetric origin. Thus our point of view at the moment is that the starting point for the search for physics beyond MSSM should start with the SUSYLR model as the first step.

## 11 Grand unification and supersymmetry

Soon after the discovery of the standard model, it became clear that embedding the model into higher local symmetries may lead to two very distinct conceptual advantages: (i) they may provide quark lepton unification [19] providing a unified understanding of the apriori separate interactions

of the two different types of matter and (ii) they can lead to description of different forces in terms of a single gauge coupling constant. How actually the unification of gauge couplings occurs was discussed in a seminal paper by Georgi, Quinn and Weinberg[20]. They used the already known fact that the coupling parameters in a theory depend on the mass scale and showed that the gauge couplings of the standard model can indeed unify at a very high scale of order  $10^{15}$  GeV or so. Although this scale might appear too far removed from the energy scales of interest in particle physics then, it was actually a blessing in disguise since in GUT theories, obliteration of the quark-lepton distinction manifests itself in the form of baryon instability such as proton decay and the rate of proton decay is inversely proportional to the 4th power of the grand unification scale and only for scales near  $10^{15}$  GeV or so, already known lower limits on proton life times could be reconciled with theory. This provided a new impetus for new experimental searches for proton decay. The minimal grand unification model based on the  $SU(5)$  group suggested by Georgi and Glashow made very precise prediction for the proton lifetime of  $\tau_p$  between  $1.6 \times 10^{30}$  yrs. to  $2.5 \times 10^{28}$  yrs. Attempts to observe proton decay at this level failed ruling out the simple minimal nonsupersymmetric  $SU(5)$  model. The primary decay mode of proton in the nonsupersymmetric  $SU(5)$  is  $p \rightarrow e^+ + \pi^0$ , which now has a lower limit on its lifetime from the Super-Kamiokande experiment[21] of  $\tau_p \geq 1.6 \times 10^{33}$  years assuming 100% branching ratio.

In fact the situation for nonsupersymmetric  $SU(5)$  was worse since it predicted a value for  $\sin^2\theta_W$  which is much lower than the experimentally observed one. This situation is depicted in Fig. 1 as a lack of unification of the gauge couplings using the experimental value of  $\sin^2\theta_W$  at the weak scale.

A revival of interest in the idea of grand unification occurred after supersymmetry became part of the phenomenology of particle physics in the early 80's. Two points were realized that led to this. First point already emphasized is that a theoretical understanding of the large hierarchy between the weak scale and the GUT scale was possible only within the framework of supersymmetry as discussed earlier. Secondly, on a more phenomenological level, measured values of  $\sin^2\theta_W$  from the accelerators coupled with the observed values for  $\alpha_{strong}$  and  $\alpha_{em}$  could be reconciled with the unification of gauge couplings only if the superpartners were included in the evolution of the gauge couplings and the supersymmetry breaking scale was assumed to be near the weak scale, which was independently motivated anyway[22](see Fig. 2).

This has led to considerations of many grand unified gauge groups, notable among them being models based on the groups  $SU(5)$  and  $SO(10)$ [23]. At present, this is an active area of research in particle theory. We do not enter into the details of this area except to remark that the discovery of a nonzero mass has put the  $SO(10)$  models at a clear advantage over the  $SU(5)$  one.

## 12 Strongly coupled supersymmetric gauge theories and composite models of quarks and leptons

As mentioned earlier in this article, supersymmetry has helped improve our understanding of the nonperturbative dynamics of nonabelian gauge theories. In particular, there has developed a paradigm concerning the nature and interaction of the low energy composite states in such models. This has revived the hope that earlier ideas concerning the possible substructure of quarks and leptons may perhaps be given a more solid field theoretical foundation and new understanding of quark lepton physics may be gleamed. It ought to be emphasized that while the above paradigm cannot be rigorously proved, it passes an impressive number of consistency tests far beyond the 't Hooft's anomaly matching condition that was the main dynamical input of the original set of composite models[24].

To illustrate the general procedure adopted in this approach, Consider a model based on the gauge group  $SU(N_c)$  with  $N_f$  flavored preons  $F$  transforming as  $N_c$  and  $F^c$  as  $N_c^*$ -dimensional representation representation of the group. If  $N_f > N_c$ , then the composite states are given by  $F_a F^{c,b}$  and  $\epsilon^{i_1 \dots i_{N_c}} F_{a_1, i_1} \dots F_{a_{N_c}, i_{N_c}}$  (plus the right chiral product). The former in the familiar QCD language are meson-like states and the latter are baryon-like states. The effective superpotential for such a composite theory will be given by allowed  $SU(N_c)$  invariant products of the preonic

fields that lead to products of above states. Thus, the dynamics of the effective theory is really quite constrained.

This paradigm has inspired a number of attempts to build composite models for quarks and leptons. However to date no compelling scenario has emerged, although the potential for model building remains good.

### 13 Supersymmetry and string theory

While in this review, we have stayed within the domain of field theories, supersymmetry plays also an essential role in the arena of string theories. To briefly describe this role, we note that string theories posit that at extremely small distances, the description of nature in terms of particles breaks down and the fundamental variables become strings embedded in spacetime. The strings were originally thought to have a size of order of the Planck length ( $\sim 10^{-33}$  cm) although recent thinkings in the subject allows much larger length scales (even upto  $10^{-18}$  cm).

Most of the interesting results in string theories derive from the property that these theories have conformal invariance which has as its subgroups scale invariance and the Lorentz invariance. Since string theories are defined in two dimensions (known as world sheet), the conformal invariance is infinite dimensional. This puts enormous number of constraints on string theories. One of the first ones is that the theory is defined only in 26 dimensions if it is pure bosonic. Also it has a tachyonic degree of freedom.

To connect it to particle physics it is postulated that the vibrational modes of the strings are to be identified with particles and the resulting Lagrangian should follow as a result of maintaining conformal invariance (i.e. setting Beta-functions to zero). It is clear that for a bosonic theory the vibrations can only be bosonic and thus cannot be realistic descriptions of nature. To generate fermions, one incorporates fermionic string degrees of freedom, same number as the bosonic degrees. ( This was where supersymmetry was first discovered in the context of particle physics.) To get fermionic particles (i.e. string states transforming as spinors under the Lorentz group), one has to impose the so called ramond boundary conditions on the fermionic strings. In any case the restrictions of conformal symmetry are now altered and one can have a situation with no tachyons and  $d=10$ . Further excitement in string theories came due to the discovery that the closed strings contain massless spin two states that can be identified with the graviton.

What is important for our discussion here is that string models provide what can be called a “derivation” of supersymmetry as a part of field theories that result in the low energy limit of field theories. Maintaining conformal invariance of string theories at the one loop level automatically require that the string theories lead to a particle spectrum which is supersymmetric. What is more: the string theories indeed lead to local supersymmetry. Also they have all the ingredients to lead to a breakdown of supersymmetry although this has not yet been demonstrated and is part of the general program of determining the correct vacuum of the string theories.

### 14 Conclusion

The advent of supersymmetry has opened up a new era of understanding and progress in both field theories and particle physics. Although so far experimentally no evidence has appeared for new particles that would signal the reality of these ideas, its impact in terms of better understanding of many aspects of physics has been so overwhelming that it has won many converts and it will indeed be a cruel irony if after so much promise, nature reveals its disapproval of this new kind of symmetry.

We have only scratched the tip of a whole mountain that supersymmetry is. Within the arena of field theories with supersymmetry, we have only discussed only  $N=1$  supersymmetry whereas there are many interesting constraints on field theories that emerge once we enlarge the number of supersymmetries to 2 to the maximum number of 8. For instance, the divergence properties of field theories keep improving as we go from  $N=1$  to  $N=2$  and finally to  $N=4$  where the field theory is a finite field theory. The improvement in  $N=2$  case is that there are infinities only



from the gauge sector but not from the Yukawa sector. For the  $N=2$  case, there is of course an exact solution discovered by Seiberg and Witten that threw new light on many of the issues in string theory as well supersymmetric field theories. Another direction in which developments have taken place is in going to higher dimensions while keeping  $N=1$ . It is known in this case that if we want to keep field theories with spin  $\leq 2$ , then the maximum allowed dimension is  $d=11$ . This 11-dimensional supergravity thought for a long time to be of mere academic interest has been found to be the low energy limit of certain string theories in the strong coupling limit and is poised to open up some very new ways to look the question of unification- for instance there may be unification of all interactions including gravity at one scale, the canonical scale of grand unification of  $2 \times 10^{16}$  GeV. Another very interesting supersymmetric theory is the  $N=1$  10-dimensional super-Yang-Mills theory coupled to gravity. This theory is known to arise in the low energy limit of the so called heterotic string theories and forms the basis of much of what is known as superstring phenomenology.

This work was supported by the National Science Foundation grant no. PHY-9802551.

## References

- [1] J Bagger, J Wess: *Supersymmetry and Supergravity*, Princeton University Press (1983).
- [2] R N Mohapatra: *Unification and Supersymmetry*, Springer-Verlag, Second edition (1991).
- [3] H Haber, G Kane: Phys. Rep. **117**, 76 (1984); R. Arnowitt, A. Chamshedine and P. Nath, *Applied  $N=1$  Supergravity*, (World Scientific, 1984).
- [4] H P Nilles: Phys. Rep. **110**, 1 (1984).
- [5] M Grisaru, M Rocek, W Siegel: Nucl. Phys. **B159**, 429 (1979).
- [6] For a review of the recent developments, see K. Intriligator and N. Seiberg, hep-th/9509066 *Proceedings of TASI95*, ed. K. T. Mahanthappa et al (World Scientific, 1995).
- [7] A. Davis, M. Dine and N. Seiberg, Phys. Lett. **125 B**, 487 (1983).
- [8] A. Nelson and M. Strassler, hep-ph/9607362; M. Luty and R. N. Mohapatra, Phys. Lett. **B396**, 161 (1997) N. Haba and N. Okamura, DPNU-97-27 (1997).
- [9] For a review and references, see P. van Nieuwenhuizen, Phys. Rep. **68**, 189 (1981).
- [10] *Supersymmetry-96: Theoretical Perspectives and Experimental Outlook*, ed. R N Mohapatra and A Rasin, North Holland (1997).
- [11] R Arnowitt, A Chamshedine, P Nath, Phys. Rev. Lett. **49**, 970 (1982); R Barbieri, S Ferrara, C Savoy, Phys. Lett. **B119**, 343 (1982).
- [12] M Dine, A Nelson: **D 48**, 1277 (1993); M. Dine, A.E. Nelson, Y. Nir, and Y. Shirman, hep-ph/9507378; Phys. Rev. **D53** (1996) 2658; A.E. Nelson, hep-ph/9511218; M. Dine and A.E. Nelson, Phys. Rev. **D48**, 1277 (1993); M. Dine, A.E. Nelson, and Y. Shirman, Phys. Rev. **D51**, 1362 (1995).
- [13] P Binetruy, E Dudas, Phys. Lett. **B389**, 503 (1996); G Dvali, A Pomarol, Phys. Rev. Lett. **77**, 3728 (1996); R N Mohapatra, A Riotto, Phys. Rev. **D55**, 4262 (1997).
- [14] M. Dine, N. Seiberg and E. Witten, Nucl. Phys. **B289**, 585 (1987); J. Attick, L. Dixon and A. Sen, *ibid* **B292**, 109 (1987).
- [15] C. S. Aulakh and R. N. Mohapatra, Phys. Lett. **119B**, 36 (1982); L. Hall and M. Suzuki, Nucl. Phys. **B231**, 419 (1984); V Barger, G F Giudice, M Y Han; Phys. Rev. **40**, 2987 (1989); For a recent review, see G Bhattacharyya, Proceedings of SUSY'96, Nucl. Phys. (Proc. Suppl.), **52A**, 83 (1996).



- [16] J. C. Pati and A. Salam, Phys. Rev. **D10**, 275 (1974); R. N. Mohapatra and J. C. Pati, Phys. Rev. **D 11**, 566, 2558 (1975); G. Senjanović and R. N. Mohapatra, Phys. Rev. **D 12**, 1502 (1975).
- [17] R. Kuchimanchi and R. N. Mohapatra, Phys. Rev. **D48**, 4352 (1993); Phys. Rev. Lett. **75**, 3989 (1995); C. Aulakh, A. Melfo and G. Senjanović, hep-ph/9707258; Z. Chacko and R. N. Mohapatra, Phys. Rev. **D 58**, 015001 (1998); C. Aulakh, K. Benakli and G. Senjanović, Phys. Rev. Lett. **79**, 2188 (1997).
- [18] M. Gell-Mann, P. Rammond and R. Slansky, in *Supergravity*, eds. D. Freedman *et al.* (North-Holland, Amsterdam, 1980); T. Yanagida, in Proc. KEK workshop, 1979 (unpublished); R.N. Mohapatra and G. Senjanović, Phys. Rev. Lett. **44**, 912 (1980).
- [19] J. C. Pati and A. Salam, Phys. Rev. **D10**, 275 (1974); H. Georgi and S. L. Glashow, Phys. Rev. Lett. **32**, 438 (1974).
- [20] H. Georgi, H. Quinn and S. Weinberg, Phys. Rev. Lett. **33**, 451 (1974).
- [21] M. Shiozawa et al. (Super-Kamiokande collaboration) ICRR-Report-419-98-15.
- [22] W. Marciano and G. Senjanović, Phys. Rev. **D25**, 3092 (1982); U. Amaldi, W. de Boer and H. Furstenuau, Phys. Lett. **B260**, 447 (1991); P. Langacker and M. Luo, Phys. Rev. **D44**, 817 (1991); J. Ellis, S. Kelly and D. Nanopoulos, Phys. Lett. **B260**, 131 (1991).
- [23] For a recent review, see R. N. Mohapatra, *TASI97-Lectures on Supersymmetry*, edited by J. Bagger (World Scientific, 1998).
- [24] J. C. Pati and A. Salam, Phys. Rev. **D10**, 275 (1973); Y. Chikashige, H. Akama and H. Terazawa, Phys. Rev. **D 15**, 480 (1977); O. W. Greenberg and J. Sucher, Phys. Lett. **B99**, 339 (1981); W. Buchmuller, R. D. Peccei and T. Yanagida, Phys. Lett. **B124**, 67 (1983); O. W. Greenberg, R. N. Mohapatra and M. Yasue, Phys. Rev. Lett. **51**, 1737 (1983); R. Barbieri, A. Masiero and G. veneziano, Phys. Lett. **B 128**, 179 (1983).

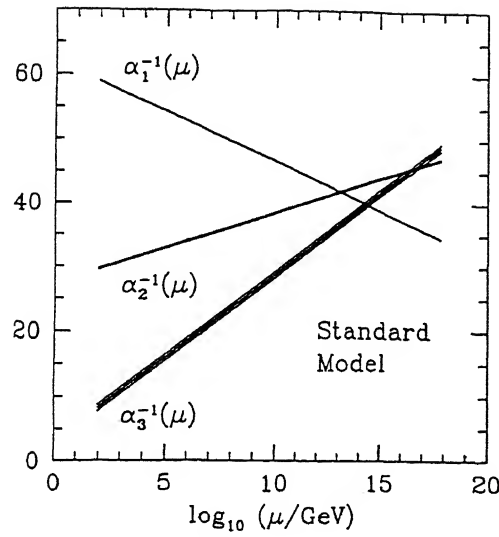


Figure 1: This figure shows the lack of unification of gauge couplings with standard model spectrum.  $\alpha_i^{-1}$  is plotted against the mass scale and the values at the weak scale are the measures values from LEP and SLC as well as other experiments. This is another way of stating the wrong prediction of  $\sin^2 2\theta_W$ .

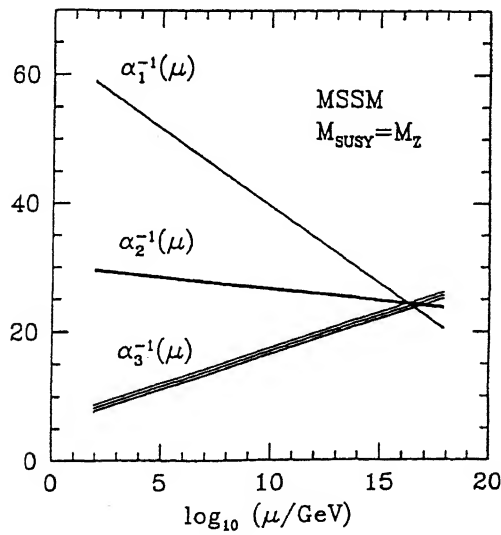


Figure 2: This figure shows the unification of gauge couplings with supersymmetric model spectrum.  $\alpha_i^{-1}$  is plotted against the mass scale and the values at the weak scale are the measures values from LEP and SLC as well as other experiments.

# 21. Supersymmetry in Field Theory

Norisuke Sakai \*

Department of Physics, Tokyo Institute of Technology

Oh-okayama, Meguro, Tokyo 152-8551, Japan

## Abstract

Supersymmetric theories are reviewed in the context of field theories. The gauge hierarchy problem in attempting the unification of all fundamental interactions is the strongest motivation of modern development of supersymmetry. Starting from the general notion of supersymmetry as a symmetry between bosons and fermions, we explain how the supersymmetry becomes a part of the space-time symmetry if we wish to maintain the relativistic invariance. The precise idea of supersymmetry is then introduced and the supersymmetric field theories are formulated. There has been a significant breakthrough in the study of nonperturbative effects in supersymmetric field theories using the holomorphy and symmetry arguments. Some of these ideas and results are briefly reviewed.

## 1 Motivations for Supersymmetry

### 1.1 Gauge Hierarchy

#### 1. Standard model

Many efforts have been devoted to study the fundamental constituents of matter and the fundamental interactions between them. At present, the experimental efforts have reached the energy scales of several 1000GeV in collisions between protons and/or antiprotons, and that of a few 100GeV in collisions between electrons and positrons.

It has been found that all the available experimental data up to these energies can be more or less adequately described by the so-called standard model. In the standard model, the fundamental constituents of matter are quarks and leptons and the three known fundamental forces in nature, strong, weak, and electromagnetic interactions are described by a gauge field theory with the  $SU(3) \times SU(2) \times U(1)$  gauge group. The standard model succeeded to describe the three fundamental interactions by a common unifying idea called the gauge principle and gave many successful predictions. The most striking confirmation of standard model is the discovery of the weak bosons,  $W$  and  $Z$  with the mass of the order of  $M_W \approx 100\text{GeV}$ . However, there are three different gauge coupling constants for each of these gauge groups  $SU(3)$ ,  $SU(2)$ , and  $U(1)$ . In that sense, the three different strengths of the three fundamental interactions are parametrized nicely, but are not quite unified. Moreover, the standard model has many input parameters that can only be determined from the experimental measurements. There are also other conceptually unsatisfactory points as well. For instance, the electric charge is found to be quantized in nature, but this phenomenon is just an accident in the standard model.

#### 2. Grand Unified Theories

Because of quantum effects, the effective gauge coupling constants change logarithmically as a function of energy scale. Then there is a possibility that the different gauge couplings

---

\*E.mail: nsakai@th.phys.titech.ac.jp

for the three fundamental interactions can become the same strength at very high energies  $M_G$ . This means that the three gauge interactions can be truly unified into a single gauge group if we choose an appropriate simple gauge group. This idea was proposed by Georgi and Glashow [1], and these models are called the grand unified theories. The grand unified theories achieved at least two good points:

- Because of simple gauge group, the electromagnetic charge is now quantized.
- Two coupling can meet at some point provided they are in the right direction. Since the grand unified theory unifies all three couplings at high energies, it gives one constraint for three couplings. Taking the two measurements of coupling constants at low energies as inputs, one can then predict the third coupling. With the simplest possibility for the unifying gauge group, this prediction was found to be not very far from the experimental data. On the other hand, the unification energy  $M_G$  is now very large compared to the mass scale  $M_W$  of the weak boson in the standard model [2]

$$\frac{M_W^2}{M_G^2} \approx \left( \frac{10^2}{10^{16}} \right)^2 \approx 10^{-28} \quad (1)$$

### 3. Gravity

Even if one does not accept the grand unified theories, one is sure to accept the existence of the fourth fundamental force, the gravitational interactions. The mass scale of the gravitational interactions is given by the Planck mass  $M_{Pl}$

$$\frac{M_W^2}{M_{Pl}^2} \approx \left( \frac{10^2}{10^{19}} \right)^2 \approx 10^{-34} \quad (2)$$

Now we have a problem of how to explain these extremely small ratios between the mass squared  $M_W^2$  to the fundamental mass squared  $M_G^2$  or  $M_{Pl}^2$  in eq.(1) or eq.(2). This problem is called the **gauge hierarchy problem**.

## 1.2 Higgs Scalar

Precisely speaking, when we say **explain** some phenomenon, we mean that it should be given a symmetry reason. This principle is called the naturalness hypothesis [3], [4]. More precisely, the system should acquire higher symmetry as we let the small parameter going to zero. The examples of the enhanced symmetry corresponding to the small mass parameter are

$$\begin{aligned} m_{J=1/2} \rightarrow 0 &\Leftrightarrow \text{Chiral symmetry} \\ m_{J=1} \rightarrow 0 &\Leftrightarrow \text{Local gauge symmetry} \end{aligned} \quad (3)$$

The mass scale  $M_W$  of weak bosons originates from the vacuum expectation value  $v$  of the Higgs scalar field. The scale of  $v$  in turn comes from the (negative) mass squared of the Higgs scalar  $\varphi$ . Therefore we need to give symmetry reasons for the extremely small Higgs scalar mass to explain the gauge hierarchy problem.

Classically the vanishing mass for scalar field does lead to an enhanced symmetry called scale invariance. However, it is well known that the scale invariance cannot be maintained quantum mechanically.

Up to now three types of possible solutions have been proposed to explain the gauge hierarchy problem.

### 1. Technicolor model

We can postulate that there is no elementary Higgs scalar at all. The Higgs scalar in the standard model has to be provided as a composite field at low energies. This option requires nonperturbative physics already at energies of the order of  $\text{TeV} = 10^3 \text{ GeV}$ . It has been

rather difficult to construct realistic models that pass all the test at low energies specially the absence of flavor-changing neutral current. Models with composite Higgs scalar are called Technicolor models [5].

## 2. Supersymmetry

Another option is to postulate a symmetry between Higgs scalar and a spinor field. Then we can postulate chiral symmetry for the spinor field to make it massless. The Higgs scalar also becomes massless because of the symmetry between the scalar and the spinor. This symmetry between scalar and spinor is called supersymmetry [6]. Supersymmetry as a possible solution of gauge hierarchy problem was proposed concretely in the context of supersymmetric grand unified theories [7] [8] [9] [10], although the use of supersymmetry has been advocated for electroweak interactions earlier [11]. Contrary to the Technicolor models, we can construct supersymmetric models that can be treated perturbatively up to extremely high energies along the spirit of the grand unified theories [12], [13].

Experimental progress for the precise measurements of coupling constants enabled one to test the unification hypothesis precisely. More than 10 years after the initial proposal of supersymmetric grand unified theories, the experimental data from LEP nicely confirmed that the nonsupersymmetric model does not give unification at a single point, and the supersymmetric model gives an excellent agreement [14].

## 3. Large extra dimensions

The most recent proposal was to note that the gravitational interactions are not tested at short distances below mm. Therefore one can consider the possibility of the fundamental scale of gravitational interactions of 1000GeV. The observed smallness of the gravitational interaction in our world is explained by imagining the extra dimensions compactified at the radius of order mm or less [15]. The supersymmetry is not needed logically in this case, although it is often used to construct concrete models.

# 1.3 Symmetry Relating Different Statistics and Spin

## 1.3.1 Symmetry Relating Different Statistics

Supersymmetry can be defined as a symmetry relating bosons and fermions. Namely particles with different statistics are related by the supersymmetry.

There is no significant constraint in formulating such a supersymmetry in nonrelativistic quantum theories. In fact the supersymmetry has been useful in several areas of nonrelativistic quantum theory such as condensed matter physics and nuclear physics. Let us mention two interesting applications:

### 1. Solid State Physics

If one considers a spin system in random magnetic fields, the randomness of the magnetic field tends to disorder the spin system. It has been found that the critical behavior of the spin system in random magnetic fields in  $d$  dimensions is the same as that of the spin system without the random magnetic fields in  $d-2$  dimensions. This phenomenon is sometimes called dimensional reduction. Parisi and Sourlas gave a beautiful explanation of this phenomenon by uncovering the underlying supersymmetry of the spin system in the random magnetic fields [16].

### 2. Nuclear Physics

In certain complex nuclei, it is quite useful to use supersymmetry among quasi particle excitations to classify various nuclear energy levels.

### 1.3.2 Symmetry Relating Different Spins

We are mainly interested in supersymmetry as a fundamental symmetry principle. We have two other fundamental principles in modern physics: quantum theory and relativity. In nature, all bosons have integer spin and all fermions have half-odd integer spin. This fact can be explained if we employ relativistic quantum field theory. Therefore supersymmetry inevitably becomes a **symmetry between particles with different spin** if we want to maintain relativistic invariance. Since the spin is a quantum number associated with the rotation, we need to formulate supersymmetry as a symmetry that is nontrivially combined with the space-time symmetry such as rotations, translations, and Lorentz transformations.

It has been a notoriously difficult problem to formulate a nontrivial symmetry that relates particles with different spins. This point can be most neatly summarized by the so-called “No-go Theorem” by Coleman and Mandula [17]. They assumed Lorentz invariance, analyticity of scattering amplitudes (corresponding to the causality), nontrivial S-matrix, and other technical assumptions. They found that Poincaré group can only appear as a direct product group with other symmetry. Namely no nontrivial symmetry is possible between particles of different spins. In this No-go theorem, they have actually assumed that all the symmetry relations are expressed in terms of commutation relations.

Much later, it has been found that nontrivial symmetry is possible if one uses anticommutation relations among symmetry generators instead of the ordinary commutation relations. With the same assumptions as those of Coleman and Mandula except the introduction of the anticommutation relation, Haag, Lopuszanski, and Sohnius were able to obtain the most general symmetry [18]. They have found that the supersymmetry as we know now is the only possible symmetry that involves space-time symmetry nontrivially. We will describe this supersymmetry in subsequent sections.

## 2 Basic Concepts in Supersymmetric Field Theory

### 2.1 Superfield and Supertransformation

To formulate symmetry such as rotation, it is most convenient to introduce a coordinate system to distinguish different directions in space. Similarly, to formulate the supersymmetry, it is useful to introduce a coordinate  $\theta$  to distinguish bosons and fermions. It has to be an anticommuting spinor, since it relates bosons and fermions. Our conventions for spinors are summarized in Appendix.A. Anticommuting number is called Grassmann number. Combined with the space-time coordinates  $x^m$ , we have  $x^m, \theta$  as coordinates in superspace.

A function  $\Phi(x, \theta)$  of  $x^m, \theta$  is called superfield. Because of anticommuting property, the superfield can be expanded in terms of Grassmann number to obtain the finite number of ordinary fields. In the case of four component Majorana spinor  $\theta$ , the superfield contains 16 component of ordinary fields. Half of them are bosons and half of them are fermions.

$$\begin{aligned} \Phi(x, \theta) = & C(x) + \bar{\theta}\psi(x) - \frac{1}{2}\bar{\theta}\theta N(x) - \frac{i}{2}\bar{\theta}\gamma_5\theta M(x) \\ & - \frac{1}{2}\bar{\theta}\gamma^m\gamma_5\theta v_m(x) + i\bar{\theta}\theta\bar{\theta}\gamma_5\lambda(x) + \frac{1}{4}(\bar{\theta}\theta)^2 D(x) \end{aligned} \quad (4)$$

Let us consider as a simplest transformation in the superspace an (infinitesimal) translation by  $\epsilon$  in the Grassmann number  $\theta$ . To make it a nontrivial space-time symmetry, we shift also the space-time coordinate as follows,

$$\delta\theta = \epsilon, \quad \delta x^m = -i\bar{\epsilon}\gamma^m\theta \quad (5)$$

This form is the simplest possibility that is Lorentz covariant and is linear in  $\epsilon$ . This transformation is called the supertransformation. With this transformation, the superfield is transformed as

$$\delta\Phi(x, \theta) = \bar{\epsilon} \left( \frac{\partial}{\partial\theta} - i\gamma^m\theta \frac{\partial}{\partial x^m} \right) \Phi(x, \theta) = - \left( \frac{\partial}{\partial\theta} - i\bar{\theta}\gamma^m \frac{\partial}{\partial x^m} \right) \epsilon \Phi(x, \theta)$$

$$\equiv [\Phi(x, \theta), \bar{\epsilon}Q] = [\Phi(x, \theta), \bar{Q}\epsilon] \quad (6)$$

The first line is represented by a differential operator in terms of the Grassmann number acting on superfield, whereas the second line is expressed as a commutator between the quantized superfields and the supercharge  $Q$  which is the unitary operator for the supersymmetry transformation. It is useful to note that the basic definition of the supertransformation dictates that the Grassmann number  $\theta, \bar{\theta}$  have dimension of the square-root of the coordinate  $x^m$ . Useful formulas for derivatives of Grassmann numbers are summarized in the Appendix.B.

To find the algebra satisfied by the supercharges, we make two successive supertransformations in eq.(5), and make the difference between the results of transformations in different order

$$\begin{aligned} & [\Phi, [\bar{\epsilon}_1 Q, \bar{Q}\epsilon_2]] = [\Phi, [\bar{\epsilon}_1 Q, \bar{\epsilon}_2 Q]] = [[\Phi, \bar{\epsilon}_1 Q], \bar{\epsilon}_2 Q] - [[\Phi, \bar{\epsilon}_2 Q], \bar{\epsilon}_1 Q] \\ &= (\delta(\epsilon_2))(\delta(\epsilon_1))\Phi - (\delta(\epsilon_1))(\delta(\epsilon_2))\Phi \\ &= \left[ \left( -\frac{\partial}{\partial\theta} + i\bar{\theta}\gamma^m\partial_m \right) \epsilon_2, \bar{\epsilon}_1 \left( \frac{\partial}{\partial\bar{\theta}} - i\gamma^n\theta\partial_n \right) \right] \Phi(x, \theta) \\ &= 2\bar{\epsilon}_1\gamma^m\epsilon_2(-i\partial_m\Phi(x, \theta)) = 2\bar{\epsilon}_1\gamma^m\epsilon_2[\Phi(x, \theta), P_m] \end{aligned} \quad (7)$$

Thus we find that the anticommutator of the supercharges is given by the space-time translation represented by the four-momentum operator  $P^m$ . This property is a direct consequence of the space-time coordinate shift bilinear in Grassmann numbers in eq.(5).

Since the chirality projection is useful in formulating supersymmetry, we shall use the two component notation for spinors from now on. The two component notation is summarized in Appendix.A. Then the anticommutators between supercharges are given by

$$\{Q_\alpha, \bar{Q}_{\dot{\beta}}\} = 2(\sigma^m)_{\alpha\dot{\beta}}P_m, \quad \{Q_\alpha, Q_\beta\} = 0, \quad \{\bar{Q}_{\dot{\alpha}}, \bar{Q}_{\dot{\beta}}\} = 0 \quad (8)$$

The translation operator  $P^m$  together with the Lorentz transformations  $J^{mn}$  form the group of space-time transformations, the Poincaré group. The other commutation relations are found to have intuitive physical meaning. First the supercharges are translation invariant and transform as a spinor under the Lorentz transformations.

$$[Q, P_m] = 0, \quad [Q_\alpha, J^{mn}] = i(\sigma^{mn})_\alpha{}^\beta Q_\beta, \quad [\bar{Q}^{\dot{\alpha}}, J^{mn}] = i(\bar{\sigma}^{mn})^{\dot{\alpha}}{}_{\dot{\beta}} \bar{Q}^{\dot{\beta}} \quad (9)$$

The rest of the algebra forms the ordinary algebra for the Poincaré group.

$$[P_m, P_n] = 0, \quad [P^m, J^{nl}] = -i(\eta^{mn}P^l - \eta^{ml}P^n) \quad (10)$$

$$[J^{mr}, J^{nl}] = -i(\eta^{rn}J^{ml} + \eta^{ml}J^{rn} - \eta^{mn}J^{rl} - \eta^{rl}J^{mn}) \quad (11)$$

Thus we find, as promised, that the supersymmetry has two characteristic features:

1. It involves the anticommutators. and
2. It is a part of spacetime symmetry.

## 2.2 Unitary Representation

Supersymmetry requires bosons and fermions to form a multiplet. To find the particle content dictated by the supersymmetry explicitly, we need to study the unitary representation of the supersymmetry algebra.

### 2.2.1 $N = 1$ Massive case

Since the supersymmetry is a part of the space-time symmetry, we should combine unitary representations of Poincaré group to form the unitary representation of the supersymmetry. To obtain

the unitary representation of the Poincaré group, we first diagonalize the four momentum  $P^m$ . For the massive case, we can choose the rest frame as the standard frame  $P^m = (M, 0, 0, 0)$ . The stability group that leaves the standard frame  $P^m = (M, 0, 0, 0)$  unchanged is the  $SO(3)$  subgroup. The unitary representation of the  $SO(3)$  subgroup is labeled by the angular momentum  $j$  and its  $z$  component  $m$ . Now we should combine these representations  $(P^m, j, m)$  of the Poincaré group to obtain the unitary representation of the supercharge  $Q$ , since  $Q$  commutes with the four momenta  $[Q, P_m] = 0$ . Since the supercharge has spin  $1/2$  as shown in eq.(9),  $Q$  changes  $j$  and  $m$  by  $\pm \frac{1}{2}$ . The anticommutators (8) between supercharges  $Q$  are precisely the same algebra as the fermion creation and annihilation operators, if we rescale by  $\sqrt{2M}$ .

$$\{Q_\alpha, Q_\beta\} = \{\bar{Q}_{\dot{\alpha}}, \bar{Q}_{\dot{\beta}}\} = 0, \quad \{Q_\alpha, \bar{Q}_{\dot{\beta}}\} = 2M\delta_{\alpha\dot{\beta}} \quad (12)$$

Since there are 2 components of spinor indices, there are 2 kinds of “fermions”. We can regard  $\bar{Q}_{\dot{\alpha}}$ ,  $\dot{\alpha} = 1, 2$  as “annihilation operators”, and  $Q_\alpha$ ,  $\alpha = 1, 2$  as “creation operators”. The unitary representations of these operators can be obtained by assuming ground state that is defined as the state annihilated by the “annihilation operators”  $\bar{Q}_{\dot{\alpha}}|j\rangle = 0$ ,  $\alpha = 1, 2$ . Here the ground state  $|j\rangle$  is assumed to be an eigenstate of angular momentum  $j$ . Since the multiplication of the same type of supercharges vanish, we obtain only four possible states by applying the “creation operator”  $Q_\alpha$

$$\begin{pmatrix} Q_1|j\rangle \\ |j\rangle \\ Q_2|j\rangle \end{pmatrix} \sim \begin{pmatrix} j - \frac{1}{2} \\ j \\ j + \frac{1}{2} \end{pmatrix} \quad (13)$$

The number of states in the multiplet is given by  $4(2j+1)$ ,  $j = 0, \frac{1}{2}, \dots$ . Two lowest multiplets of the massive supermultiplet are explicitly shown in the table.

1.  $j = 0$  case  $\Rightarrow$  Chiral scalar multiplet

spin $j$	field	degree of freedom
0	two real scalar	2
1/2	a Majorana spinor	2

2.  $j = \frac{1}{2}$  case  $\Rightarrow$  Vector multiplet

spin $j$	field	degree of freedom
0	a real scalar	1
1/2	2 Majorana spinor	4
1	a real vector	3

### 2.2.2 $N = 1$ Massless case

In the case of massless particles, we can choose the standard frame as  $P^m = (P, 0, 0, P)$ . The stability group that leaves the standard frame  $P^m = (P, 0, 0, P)$  unchanged is the Euclid group in two dimensions  $E_2$ :  $E_2 = (J^{12}, J^{01} - J^{31}, J^{02} + J^{23})$ . It is well-known that the unitary representation of massless particles is labeled by the helicity  $J^{12}$  [19]. In the standard frame, the nonvanishing anticommutator between supercharges is given by

$$\{Q_\alpha, \bar{Q}_{\dot{\beta}}\} = 2(\sigma_0 + \sigma_3)_{\alpha\dot{\beta}} P = 4P \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad (14)$$

$$\{Q_1, \bar{Q}_1\} = 4P, \quad \bar{Q}_1 = (Q_1)^* \quad (15)$$

Therefore we have only single fermion “creation and annihilation operators”. If we take the state of helicity  $\lambda$  as the ground state  $\bar{Q}_1|\lambda\rangle = 0$ , we obtain a multiplet consisting of only 2 states whose helicities differ by  $1/2$ .

$$(|\lambda\rangle, Q_1|\lambda\rangle) \sim \left(|\lambda\rangle, \left|\lambda - \frac{1}{2}\right\rangle\right) \quad (16)$$



Although the number of states in a multiplet is two, it is often required that the CPT invariance necessitates to combine states with opposite helicity if they are not in the same multiplet. Then the number of states becomes four. Frequently used multiplets are shown in the table.

$$(\lambda, \lambda - \frac{1}{2}, -\lambda + \frac{1}{2}, -\lambda) \quad (17)$$

highest helicity	helicities of fields	name of multiplet
$\lambda = \frac{1}{2}$	$(\frac{1}{2}, 0, 0, -\frac{1}{2})$	chiral scalar multiplet
$\lambda = 1$	$(1, \frac{1}{2}, -\frac{1}{2}, -1)$	vector multiplet
$\lambda = 2$	$(2, \frac{3}{2}, -\frac{3}{2}, -2)$	graviton multiplet

### 2.2.3 Extended Supersymmetry

The most general supersymmetry algebra found by Haag et. al. contains  $N$  species of supercharges  $Q^L$  [18]. It is called the  $N$ -extended supersymmetry. In two component notation, it reads

$$\{Q_\alpha^L, \bar{Q}_{\dot{\beta}M}\} = 2(\sigma^m)_{\alpha\dot{\beta}} P_m \delta_M^L, \quad \{Q_\alpha^L, Q_\beta^M\} = \epsilon_{\alpha\beta} X^{LM}, \quad \{\bar{Q}_{\dot{\alpha}L}, \bar{Q}_{\dot{\beta}M}\} = \epsilon_{\dot{\alpha}\dot{\beta}} X_{LM}^\dagger \quad (18)$$

$$[X^{KL}, Q_\alpha^M] = [X^{KL}, \bar{Q}_{\dot{\alpha}M}] = [X^{LM}, X^{KN}] = 0 \quad (19)$$

where  $X$  are called central charges.

1. Let us first consider the massless case without central charges  $X$ . Similarly to eq.(14), the  $N$ -extended supersymmetry gives  $Q_1^L$ ,  $L = 1, \dots, N$  as fermion "creation operators", if there are no central charges. Starting from the ground state with the helicity  $\lambda$ , we descend in helicity by half unit in each step by operating  $Q_1^L$ .

$$\begin{array}{ccccccc} |\lambda\rangle & \rightarrow & |\lambda - \frac{1}{2}\rangle & \rightarrow & |\lambda - 1\rangle & \cdots & \rightarrow & |\lambda - \frac{N}{2}\rangle \\ 1 & & N & & \left(\begin{array}{c} N \\ 2 \end{array}\right) & \cdots & & 1 \end{array} \quad (20)$$

The number of states is denoted below each helicity states and sums to  $2^N$ . If the multiplet is not CPT self-conjugate, CPT conjugate states should be added. Two points are worth mentioning:

- There are a number of arguments suggesting that consistent formulation of interacting massless fields is limited to spin up to two in four-dimensions. This limits the highest helicity to be less than or equal to 2.

$$|\lambda| \leq 2, \quad |\lambda - \frac{N}{2}| \leq 2 \quad (21)$$

Therefore the highest possible supersymmetry is  $N = 8$  which gives  $4 \times 8 = 32$  supercharges. The  $N = 8$  supersymmetry in four-dimension is maximal, and it automatically contains graviton ( $\lambda = \pm 2$ ). Therefore the interacting  $N = 8$  supersymmetric theory is nothing but the  $N = 8$  extended supergravity.

- If one wants a renormalizable theory, highest helicity should be one or less. This limits  $N$  to be less than or equal to  $N = 4$ :  $J \leq 1 \Rightarrow N \leq 4$ . The maximal case gives the maximally supersymmetric gauge theory:  $N = 4$  supersymmetric Yang-Mills theory.

2. Massive  $N$ -extended supersymmetry case without central charge  $X$  allows  $2N$  supercharges  $Q_1^L, Q_2^L, \dots, Q_N^L$ ,  $L = 1, \dots, N$  as fermion “creation operators”. We thus obtain the number of states in a multiplet to be  $2^{2N}(2j+1)$ ,  $j = 0, \frac{1}{2}, \dots$ .
3. BPS states:

If we have massive  $N$ -extended supersymmetry case with central charge  $X$ , we can have interesting situation called the BPS states where only a part of supersymmetry is maintained giving the smaller number of states in a single multiplet. Such a multiplet is sometimes called a short representation.

Let us illustrate by an example in the  $N = 2$  case that has  $SU(2)$  as an internal symmetry. Since the central charge has to be proportional to the invariant tensor of the internal symmetry  $SU(2)$ , we parametrize

$$X^{LM} = 2Z\epsilon^{LM} \quad (22)$$

Let us take the rest frame  $P^m = (M, 0, 0, 0)$ . The  $N = 2$  supersymmetry algebra becomes

$$\{Q_\alpha^L, \bar{Q}_{\dot{\beta}M}\} = 2M\delta_{\alpha\dot{\beta}}\delta_M^L, \quad \{Q_\alpha^L, Q_\beta^M\} = 2Z\epsilon_{\alpha\beta}\epsilon^{LM}, \quad \{\bar{Q}_{\dot{\alpha}L}, \bar{Q}_{\dot{\beta}M}\} = 2Z^*\epsilon_{\dot{\alpha}\dot{\beta}}\epsilon_{LM} \quad (23)$$

Then we have to consider both chirality of supercharges together. Since the anticommutator matrix must be positive definite matrix, we obtain an inequality

$$M \geq |Z| \quad (24)$$

This bound is called the BPS (Bogomolnyi-Prasad-Sommerfield) bound [20].

If  $M = |Z|$ , there are zero eigenvalues for the matrix. This implies that a linear combination of  $Q$ 's annihilates all the states, and cannot be used to create physical states. Therefore we obtain smaller number of particle states to represent the supersymmetry algebra. For example, if we have  $Z = M$ , we find convenient linear combinations of supercharges as

$$Q^{(1)} \equiv Q_1^{L=1} + \bar{Q}_{\dot{2}L=2}, \quad Q^{(2)} \equiv Q_2^{L=1} - \bar{Q}_{\dot{1}L=2}, \quad (25)$$

These satisfy

$$\{Q^{(i)}, Q^{(j)\dagger}\} = 4M\delta^{ij} \quad (26)$$

and all other anticommutators vanish.

This is algebraically the same as the case of the massive  $N = 1$  supersymmetry. Therefore the number of states is reduced by  $1/4$ :  $2^4(2j+1) \rightarrow 2^2(2j+1)$ .

This phenomenon occurs when the determinant of the anticommutators of supercharges vanishes. The resulting multiplet contains a smaller number of physical states and is called the BPS saturated states [21].

The physical origin of the central charge is often given by various nonperturbative objects such as monopoles, dyons, domain walls, in general some kind of solitons.

## 2.3 Field Theory Realization

### 2.3.1 Irreducible Representation

The smallest unitary representation of the  $N = 1$  supersymmetry in four space-time dimensions requires two real spin 0 particles and two spin  $1/2$  particles. On the other hand, the general superfield  $\Phi(x, \theta, \bar{\theta})$  has 8 boson fields and 8 fermion fields, as we have seen in eq.(4).

To obtain smaller number of components than the general superfield, we should find a constraint that is compatible with the supersymmetry transformation to realize the supersymmetry in a smaller space. This is a key ingredient to construct supersymmetric field theories.

We note that the general spinors  $\theta_\alpha$  in four space-time dimensions has four components, whereas the chirally projected spinors  $\theta_\alpha, \bar{\theta}_{\dot{\alpha}}$  have only two components. Therefore if we can construct a

superfield that depends only on the chirally projected spinors, we should be able to reduce the number of component fields to half of those of the general superfield. Therefore we are tempted to use the constraint that the superfield be independent of the Grassmann number with one of the chirality

$$\frac{\partial}{\partial \bar{\theta}^{\dot{\alpha}}} \Phi(x, \theta, \bar{\theta}) = 0 \quad (27)$$

Unfortunately even if this constraint is imposed, it is not satisfied after the supersymmetry transformation.

$$\left\{ \frac{\partial}{\partial \bar{\theta}^{\dot{\alpha}}}, Q_{\beta} \right\} \neq 0 \quad (28)$$

Therefore this constraint is not consistent with supersymmetry. We can modify the derivative with respect to the Grassmann number by an additional term. We define the following covariant derivatives

$$\bar{D}_{\dot{\alpha}} \equiv -\frac{\partial}{\partial \bar{\theta}^{\dot{\alpha}}} - i\theta^{\alpha} \sigma_{\alpha\dot{\alpha}}^m \partial_m \quad (29)$$

$$D_{\alpha} \equiv \frac{\partial}{\partial \theta^{\alpha}} + i\sigma_{\alpha\dot{\alpha}}^m \bar{\theta}^{\dot{\alpha}} \partial_m \quad (30)$$

These covariant derivatives anticommute with the supersymmetry transformation

$$\{D_{\alpha}, Q_{\beta}\} = \{D_{\alpha}, \bar{Q}_{\dot{\beta}}\} = \{\bar{D}_{\dot{\alpha}}, Q_{\beta}\} = \{\bar{D}_{\dot{\alpha}}, \bar{Q}_{\dot{\beta}}\} = 0 \quad (31)$$

Therefore they can be used to constrain the superfield to reduce the number of component fields by half.

$D_{\alpha}, \bar{D}_{\dot{\alpha}}$  satisfy the same algebra as  $Q_{\alpha}, \bar{Q}_{\dot{\alpha}}$ .

$$\{D_{\alpha}, \bar{D}_{\dot{\alpha}}\} = -2i\sigma_{\alpha\dot{\alpha}}^m \partial_m, \quad \{D_{\alpha}, \bar{D}_{\dot{\beta}}\} = \{D_{\dot{\alpha}}, \bar{D}_{\dot{\beta}}\} = 0 \quad (32)$$

### 2.3.2 Chiral Scalar Field

By using the covariant derivatives, we can now define the superfield which has half as many components as the general superfields in eq.(4). Since the supercharge anticommute with the covariant derivative as shown in eq.(31), these chiral scalar fields can be used as a representation space of supersymmetry.

The (negative) chiral scalar superfield is defined by

$$\bar{D}_{\dot{\alpha}} \Phi(x, \theta, \bar{\theta}) = 0 \quad (33)$$

We can easily see that the following combination of variables satisfies this constraint

$$y^m \equiv x^m + i\theta\sigma^m\bar{\theta}, \quad \bar{D}_{\dot{\alpha}} y^m = 0 \quad (34)$$

Therefore the general solution of the constraint is simply that the superfield depends on the  $\bar{\theta}$  only through the combination  $y^m \equiv x^m + i\theta\sigma^m\bar{\theta}$ .

$$\Phi(y, \theta) = A(y) + \sqrt{2}\theta\psi(y) + \theta\theta F(y) \quad (35)$$

The supertransformation of the chiral scalar superfield is given by means of the derivative operator defined in eq.(6). In the two component notation, we obtain

$$\delta_{\xi} \Phi(y, \theta) = \left[ \xi^{\alpha} \left( \frac{\partial}{\partial \theta^{\alpha}} - i\sigma_{\alpha\dot{\alpha}}^m \bar{\theta}^{\dot{\alpha}} \frac{\partial}{\partial x^m} \right) + \left( -\frac{\partial}{\partial \bar{\theta}^{\dot{\alpha}}} + i\theta^{\alpha} \sigma_{\alpha\dot{\alpha}}^m \frac{\partial}{\partial x^m} \right) \bar{\xi}^{\dot{\alpha}} \right] \Phi(x, \theta) \quad (36)$$

In terms of the component fields, we find

$$\delta_{\xi} A = \sqrt{2}\xi\psi \quad (37)$$

$$\delta_{\xi} \psi = i\sqrt{2}\sigma^m \bar{\xi} \partial_m A + \sqrt{2}F \quad (38)$$

$$\delta_\xi F = i\sqrt{2}\bar{\xi}\bar{\sigma}^m\partial_m\psi \quad (39)$$

It is important to note that the last component  $F$  is transformed into a derivative of the lower component. The supertransformation should increase the mass dimension by  $M^{\frac{1}{2}}$ . However, the last component has the highest mass dimension and there is no component fields available except to consider the derivative of the lower component fields. This point is always true for the last component of the superfields. Hence the last component of the general superfield also transforms into a total derivative of lower component fields.

It is important to realize that the chiral scalar field is complex. Therefore the scalar component  $A$  is a complex scalar field, and the fermionic component  $\psi$  is a complex Weyl spinor. Let us count the number of the degrees of freedom of component fields. If we do not use the equation of motion, there are two real scalar components from  $A$  and two real scalar components from  $F$ , and four real fermionic components from  $\psi$ . We call this situation off-shell. Later we will construct the Lagrangian for this chiral scalar field. There we will find that  $\psi$  obeys the Dirac equation which reduces the on-shell degrees of freedom to half. Namely we have only a left-handed fermion and its anti-particle. As we noted previously, the mass dimension of the Grassmann number  $\theta, \bar{\theta}$  is  $M^{-\frac{1}{2}}$ . Therefore the mass dimension of the field  $F$  is  $M^2$  if we take the mass dimension of the scalar component  $A$  to be  $M^1$  as ordinarily required for the scalar field. As we will find when constructing the Lagrangian, this implies that the  $F$  cannot have ordinary kinetic term with two derivatives and is an auxiliary field that can be expressed in terms of other fields. We summarize the counting of the number of degrees of freedom in the table.

fields	real or complex spin	off-shell real d.o.f.	on-shell real d.o.f.
$A$	complex scalar	2	2
$\psi$	complex 2-comp. spinor	4	2
$F$	complex aux. scalar	2	0

Similarly, we can define the positive chiral scalar superfield by

$$D_\alpha\Phi(x, \theta, \bar{\theta}) = 0 \quad (40)$$

The general solution of the constraint is given by

$$\Phi(y^*, \bar{\theta}) = A^*(y^*) + \sqrt{2}\bar{\theta}\bar{\psi}(y^*) + \bar{\theta}\bar{\theta}F^*(y^*) \quad (41)$$

Clearly the product of chiral scalar superfields is still a chiral scalar superfield as long as the chirality is the same. On the other hand, the product of positive chiral and negative chiral scalar fields is a general superfield (without a definite chirality).

The complex conjugation changes the chirality, since the complex variable  $y^m$  is changed into  $(y^m)^*$  and the chirality of spinor is also changed by the complex conjugation  $(\theta)^* = \bar{\theta}$

$$(\Phi(y, \theta))^* = A^*(y^*) + \sqrt{2}\bar{\theta}\bar{\psi}(y^*) + \bar{\theta}\bar{\theta}F^*(y^*) \quad (42)$$

### 2.3.3 Lagrangian Field Theory with Chiral Scalar Fields

As we noted in sect.2.3.2, the last components of superfields always transform into a total derivative. There are two possibilities for the superfields: chiral scalar superfield and general superfield. Therefore we have two candidates for the Lagrangian invariant under supersymmetry transformation up to a total divergence:

1.  $D$ -term of general superfield  $\Phi$  in eq.(4)

$$[\Phi(\theta, \bar{\theta})]_D = \frac{1}{4}D^2\bar{D}^2\Phi(\theta, \bar{\theta}) \quad (43)$$

Since the product of chiral scalar superfield with opposite chirality is a general superfield, we can take the  $D$  term of the product.

2.  $F_{\pm}$ -term of chiral scalar superfield  $\Phi(\theta), \bar{\Phi}(\bar{\theta})$

$$[\Phi]_F = \frac{1}{2}D^2\Phi, \quad [\bar{\Phi}]_{\bar{F}} = \frac{1}{2}(\bar{D})^2\bar{\Phi} \quad (44)$$

Let us consider Lagrangian field theory consisting of chiral scalar fields. Since the supertransformation does not leave any product of chiral scalar fields invariant, we have to be satisfied with the invariance up to total divergence.

It is quite useful to examine the dimensions of various fields. To give the canonical dimension to the scalar component  $[A] = M^1$ , we usually assume the dimension of the chiral scalar fields to be  $M^1$ .

$$[\Phi(\theta)] = [\Phi(\bar{\theta})] = M^1 \quad (45)$$

Since the mass dimensions of the Grassmann number is half of that of the coordinates,

$$[\theta] = [\bar{\theta}] = L^{\frac{1}{2}} = M^{-\frac{1}{2}}, \quad (46)$$

we obtain that the covariant derivative has the mass dimensions as  $M^{\frac{1}{2}}$

$$[D] = [\bar{D}] = M^{\frac{1}{2}} \quad (47)$$

A renormalizable Lagrangian in four space-time dimensions requires that the Lagrangian should consist of operators with dimension  $\leq 4$ . We can list possible terms as follows.

1. D-type:

$$\bar{D}^2 D^2 \Phi \bar{\Phi} \quad (48)$$

Since the mass dimension of the product of covariant derivatives is  $[\bar{D}^2 D^2] = M^2$ , we see that there are no terms of this class.

2. F-type:

$$D^2(a\Phi_1 + b\Phi_1\Phi_2 + c\Phi_1\Phi_2\Phi_3) = D^2 P(\Phi) \quad (49)$$

Since  $D^2$  has dimension  $M^1$ , up to third order polynomials of chiral scalar superfields of one chirality are renormalizable. To maintain the hermiticity of the Lagrangian, we need to add hermitian conjugate terms which consist of the chiral scalar fields of opposite chirality with conjugate coefficients. The polynomial of chiral scalar superfield of the same chirality is called superpotential  $P$ .

Now let us illustrate the above consideration with a simple example: general Lagrangian with a single chiral scalar field

$$L = L_{\text{kin}} + L_{\text{int}}. \quad (50)$$

$$\begin{aligned} L_{\text{kin}} &= \frac{1}{4}D^2\bar{D}^2\Phi^*\Phi \\ &= \frac{1}{4}\partial^2 A^*A - \frac{1}{2}\partial_\nu A^*\partial^\nu A + \frac{1}{4}A^*\partial^2 A \\ &+ F^*F + \frac{1}{2}i\bar{\psi}\bar{\sigma}^\mu\partial_\mu\psi - \frac{1}{2}i\partial_\mu\bar{\psi}\bar{\sigma}^\mu\psi \\ &= -\partial_\nu A^*\partial^\nu A - i\partial_\mu\bar{\psi}\bar{\sigma}^\mu\psi + F^*F + \text{total derivatives} \end{aligned} \quad (51)$$

$$\begin{aligned} L_{\text{int}} &= \frac{1}{4}D^2\left(\frac{1}{3}f\Phi^3 + \frac{m}{2}\Phi^2 + \text{h.c.} + s\Phi\right) \\ &= f(FA^2 - \psi\psi A) + m\left(FA - \frac{1}{2}\psi\psi\right) + sF + \text{h.c.} \end{aligned} \quad (52)$$

The Euler-Lagrange equation for  $F$  is given by

$$F^* + fA^2 + mA + s = 0 \quad (53)$$

By solving this equation, we can eliminate the auxiliary field  $F$  from the Lagrangian  $L$

$$\begin{aligned} L \rightarrow & -\partial_\nu A^* \partial^\nu A + \frac{1}{2} \bar{\psi} i \bar{\sigma}^\mu \partial_\mu \psi - \frac{m}{2} \bar{\psi} \psi \\ & - (f \psi \psi A^* + \text{h.c.}) - |fA^2 + mA + s|^2 \end{aligned} \quad (54)$$

Let us suppose temporarily that the vacuum expectation value of the scalar field  $A$  vanishes. Then the parameter  $m$  gives the mass of a Majorana spinor  $\psi$  and a complex scalar  $A$ . The parameter  $f$  gives the Yukawa coupling and the scalar four point coupling  $|A^2|^2$  in the potential.

### 2.3.4 Supersymmetric gauge theory

Ordinary local gauge transformation for the matter field  $\psi(x)$  in the representation corresponding to a matrix  $T^a$  is given by

$$\psi(x) \rightarrow e^{-i\Lambda^a(x)T^a} \psi(x) \quad (55)$$

The matter field should be extended to a chiral scalar superfield  $\Phi(x, \theta)$  in the supersymmetric theory. In order to maintain chirality of the superfield, we need to extend the gauge parameter function  $\Lambda(x)$  to be a chiral scalar superfield  $\Lambda(x, \theta)$ .

$$\Phi(x, \theta) \rightarrow \exp(-i\Lambda^a(x, \theta)T^a) \Phi(x, \theta) \quad (56)$$

Since the chiral scalar superfield contains a complex scalar field, supersymmetrized local gauge transformation actually contains scale transformations.

The kinetic term of the matter fields should be made gauge invariant by introducing the gauge field. In supersymmetric field theory, the kinetic term of the chiral scalar fields consists of product of chiral scalar field with opposite chirality  $\Phi^* \Phi$  as in eq.(51). Therefore we need to introduce a general superfield as in eq.(4) instead of chiral scalar superfield. We see immediately that the general superfield contains vector field as a component. For this reason, the general superfield is sometimes called the vector superfield. The vector superfield  $V$  can be expanded in terms of  $\theta, \bar{\theta}$  to obtain component fields

$$\begin{aligned} V(x, \theta) \equiv & C(x) + i\theta\chi(x) - i\bar{\theta}\bar{\chi}(x) \\ & + \frac{i}{2}\theta\theta(M + iN) - \frac{i}{2}\bar{\theta}\bar{\theta}(M - iN) - \theta\sigma^m\bar{\theta}v_m(x) + i\theta\theta\bar{\theta}(\bar{\lambda}(x) + \frac{i}{2}\bar{\sigma}^m\partial_m\chi(x)) \\ & - i\bar{\theta}\bar{\theta}\theta(\lambda(x) + \frac{i}{2}\sigma^m\partial_m\bar{\chi}(x)) + \frac{1}{2}\theta\theta\bar{\theta}\bar{\theta}\left(D(x) + \frac{1}{2}\partial^2 C(x)\right) \end{aligned} \quad (57)$$

With this vector superfield, the supersymmetric version of the gauge transformations is given by

$$e^{2gV} \rightarrow e^{-i\Lambda^\dagger} e^{2gV} e^{i\Lambda} \quad (58)$$

Here the general superfield  $V \equiv V^a T^a$  belongs to the adjoint representation of the gauge group and  $g$  is the gauge coupling constant. It is dimensionless and real.

$$V^{a*} = V^a \quad (59)$$

With this gauge transformation, the kinetic term for the chiral scalar superfield becomes gauge invariant.

$$\text{tr}(\bar{\Phi} e^{2gV} \Phi) \rightarrow \text{tr}(\bar{\Phi} e^{2gV} \Phi) \quad (60)$$

In order to examine the gauge transformation of the vector supermultiplet, it is simplest if we consider the  $U(1)$  case

$$V \rightarrow V + \frac{i}{2g}(\Lambda - \Lambda^*) \quad (61)$$

$$\Lambda = A + \sqrt{2}\theta\psi + \theta\theta F \quad (62)$$

$v^m$  is an ordinary gauge field with  $\text{Re}A$  as the ordinary (real) gauge transformation parameter

$$v^m \rightarrow v^m + \frac{1}{2g}\partial^m(A + A^*) \quad (63)$$

$\lambda, D$  are gauge invariant.

$$\begin{aligned} \lambda &\rightarrow \lambda \\ D &\rightarrow D \end{aligned} \quad (64)$$

$C, \chi, M, N$  can be gauged away by  $\text{Im}A, \psi, F$  in supersymmetric gauge parameter superfield  $\Lambda$

$$\begin{aligned} C &\rightarrow C + \frac{i}{2g}(A - A^*) \\ \chi &\rightarrow \chi + \sqrt{2}\frac{1}{2g}\psi \\ M + iN &\rightarrow M + iN + \frac{1}{2g}F \end{aligned} \quad (65)$$

By exploiting the supersymmetric version of the gauge transformation, we can go to the Wess-Zumino gauge that is most popular to unravel the physical particle content of the model.

$$\begin{aligned} V_{WZ} &= -\theta\sigma^m\bar{\theta}v_m(x) + i\theta\theta\bar{\theta}\bar{\lambda}(x) \\ &\quad - i\bar{\theta}\bar{\theta}\theta\lambda(x) + \frac{1}{2}\theta\theta\bar{\theta}\bar{\theta}D(x) \end{aligned} \quad (66)$$

Since we have used the gauge transformation to go to the Wess-Zumino gauge, the Wess-Zumino gauge is not manifestly supersymmetric. In this gauge, supersymmetry is no longer manifest, but the invariance under the ordinary gauge transformation remains. The particle content can be most easily seen in the Wess-Zumino gauge.

To form a Lagrangian, we need to build the gauge field strength as a gauge covariant building block. Among component fields of the vector superfield  $V$ , the gaugino field  $\lambda^a(x)$  is the gauge covariant field with lowest dimension. We can obtain this component by applying the covariant derivative  $D$  once and  $\bar{D}$  twice.

$$W_\alpha \equiv \frac{-1}{8g}(\bar{D}\bar{D})\left(e^{-2gV^aT^a}D_\alpha e^{2gV^aT^a}\right) = -i\lambda_\alpha + \dots \quad (67)$$

Since we have differentiated twice in  $\bar{\theta}$ ,  $W_\alpha$  is a negative chiral superfield and gauge covariant

$$\bar{D}_{\dot{\beta}}W_\alpha = 0 \quad (68)$$

$$W_\alpha \rightarrow e^{-i\Lambda^aT^a}W_\alpha e^{i\Lambda^aT^a} \quad (69)$$

Similarly a positive chiral field strength is given by

$$\bar{W}_{\dot{\alpha}} = \frac{-1}{8g}(DD)(e^{2gV^aT^a}\bar{D}_{\dot{\alpha}}e^{-2gV^aT^a}) \quad (70)$$

Since supersymmetric gauge field strength is a chiral superfield, the kinetic term for vector superfield is given by the  $F$  term of the square of the supersymmetric field strength

$$L_{\text{gauge}} = \frac{1}{8}D^2(W^\alpha W_\alpha) + \text{h.c.} \quad (71)$$

In the Wess-Zumino gauge, the Lagrangian is given in terms of the component fields as

$$L_{\text{gauge}} = i\bar{\lambda}\sigma^m\partial_m\lambda - \frac{1}{4}v_{\mu\nu}^a v^{a\mu\nu} + \frac{1}{2}D^a D^a \quad (72)$$

$$v_{\mu\nu} = \partial_\mu v_\nu - \partial_\nu v_\mu + ig[v_\mu, v_\nu] \quad (73)$$

$$\nabla_\mu\lambda = \partial_\mu\lambda + ig[v_\mu, \lambda] \quad (74)$$

Similarly to the  $F$  fields, the last component  $D^a$  is an auxiliary field.

## 2.4 The General $N = 1$ Supersymmetry Lagrangian up to two Derivatives

Since we are interested in effective action, we should not require the action to be renormalizable. Here we will write down the most general  $N = 1$  supersymmetry Lagrangian in flat space (without gravity) which has up to two derivatives of fields. We have the following building blocks

1. Field content

Chiral superfield  $\Phi$

Vector superfield  $V$

2. Superpotential  $P(\Phi)$

Interaction between chiral scalar fields are given by a function called superpotential which depends on the chiral scalar fields of the same chirality only.

3. Kähler potential  $K(\Phi^\dagger, \Phi)$

The kinetic term of the chiral scalar superfields is given by the  $D$  term of a general superfield that is given by a function of chiral scalar superfields of both chirality. Since the  $D$  term is taken, the kinetic term of the action is unchanged by a transformation with a function  $f(\Phi)$  and  $\bar{f}(\bar{\Phi})$

$$K(\Phi^\dagger, \Phi) \rightarrow K(\Phi^\dagger, \Phi) + f(\Phi) + \bar{f}(\bar{\Phi}) \quad (75)$$

This invariance is called Kähler invariance. This function can be regarded as giving a geometry of field space of the chiral scalar superfields. This geometry is called Kähler metric and the function is called Kähler potential. Additional term due to the gauge interaction is denoted as  $\Gamma$ .

4. Gauge kinetic function  $H_{ab}(\Phi)$

Since the gauge kinetic term is given by the  $F$  term of supersymmetric gauge field strength, it can be multiplied by a function of chiral scalar fields which is called the gauge kinetic function.

5. Fayet-Iliopoulos D-term for  $U(1)$   $\xi$

Since  $U(1)$  vector superfield is neutral, the  $D$  term of the vector superfield is neutral and transforms into total derivative under the supertransformation. Therefore one can add a  $D$  term of the  $U(1)$  vector superfield itself  $V^{(1)}$  into the Lagrangian.

We shall denote the  $F$ -type term as  $|_{\theta\theta}$  or  $|_{\bar{\theta}\bar{\theta}}$  and  $D$ -type term as  $|_{\theta\theta\bar{\theta}\bar{\theta}}$

$$\begin{aligned} \mathcal{L} = & [K(\Phi^\dagger, \Phi) + \Gamma(\Phi^\dagger, \Phi; V)]|_{\theta\theta\bar{\theta}\bar{\theta}} \\ & + \left( \frac{1}{4} H_{ab}(\Phi) W^{a\alpha} W_\alpha^b \right)|_{\theta\theta} + h.c. + 2\xi V^{(1)}|_{\theta\theta\bar{\theta}\bar{\theta}} \\ & + (P(\Phi)|_{\theta\theta} + h.c.) \end{aligned} \quad (76)$$

The minimal forms of the Kähler potential and gauge kinetic function are given by

$$K + \Gamma = \Phi^\dagger e^{2V} \Phi, \quad (77)$$

$$H_{ab} = \frac{1}{g^2} \delta_{ab} \quad (78)$$

On the other hand, an interesting example of the nonminimal gauge kinetic function is given by

$$H_{ab}(S) = \frac{1}{g^2} \delta_{ab} + S \delta_{ab} + \dots \quad (79)$$



where  $S$  is a chiral scalar superfield which is a singlet of the gauge group. The mass dimension of the chiral scalar superfield and the superpotential  $P$  is

$$[\Phi] = M, \quad [P(\Phi)] = M^3 \quad (80)$$

If renormalizability is required, the superpotential  $P(\Phi)$  should be cubic or less in  $\Phi$ .

The equation of motion for the auxiliary field  $F^{j*}$  is given by

$$g_{ij*} F^i - \frac{1}{2} g_{kj*} \Gamma_{ml}^k \chi^m \chi^l + \frac{\partial P^*}{\partial A^{*j}} = 0 \quad (81)$$

$$g_{ij*} = \frac{\partial^2 K}{\partial A^i \partial A^{*j}} \quad (82)$$

$$\Gamma_{ml}^k = g^{kn*} \frac{\partial}{\partial A^l} g_{mn*} = g^{kn*} \frac{\partial^3 K}{\partial A^l \partial A^m \partial A^{*n}} \quad (83)$$

The equation of motion for auxiliary field  $D$  for minimal kinetic term is given by

$$\frac{1}{g} D^a + \Sigma_k A^{*k} T^a A^k = 0 \quad (84)$$

$$\frac{1}{e} D + \Sigma_k A^{*k} Q A^k + \xi = 0 \quad (85)$$

## 2.5 Perturbative Nonrenormalization Theorem

It has been very useful to use superfield perturbation theory to organize the perturbative corrections. The most interesting prediction of the superfield perturbation was the nonrenormalization theorems [22] [23]. Since the interaction among chiral scalar superfield consists of superfields with the same chirality, there is a selection rule based on purely algebraic identities on the chirality structure of possible loop corrections. By performing the algebra of Grassmann numbers, it has been shown that the loop corrections to all orders of perturbation do not give any  $F$ -type terms. This implies that not only the divergent terms but also finite terms do not appear in the  $F$ -type terms. The loop corrections in quantum effects appear only in the  $D$ -type terms. Therefore the following local terms can be generated in quantum effects.

### 1. Kähler potential $K(\Phi, \bar{\Phi})$

This gives the kinetic term of chiral scalar multiplet

### 2. Gauge kinetic function $H_{ab}(\Phi)$

This can give the nonminimal kinetic term for vector multiplet. Although the gauge kinetic term is written as a  $F$ -type term, the gauge field strength actually involves the covariant derivative of opposite chirality. Therefore it can be generated in loop corrections.

### 3. Fayet-Iliopoulos D-term for $U(1)$

As a consequence, we obtain the following

#### 1. No quadratic divergences

Typically the mass parameter can get quadratic divergences, but there is no loop corrections at all for parameters appearing in superpotential such as the mass parameters.

#### 2. No quantum corrections to masses and Yukawa couplings

For such parameters in the superpotential, even a finite correction is absent.

#### 3. Only wave function renormalization and gauge coupling renormalization are needed.

They are typically logarithmically divergent.

Let us emphasize that the necessity of the wave function renormalization means that the parameters such as mass, Yukawa coupling constant still run as one changes the scale. Therefore it is still meaningful to consider these parameters as effective coupling constants that depend on the energy scale. It should also be stressed that the above nonrenormalization theorem is obtained by the perturbation theory and is valid to all orders of perturbation. Therefore the nonperturbative effects can violate the nonrenormalization theorem.

Another interesting perturbative result is that the beta function is exactly given by 1-loop in the  $N = 2$  supersymmetric gauge theories [24].

## 2.6 R-symmetry

In supersymmetric theories, one can define a new type of symmetry called the R-symmetry. This is a continuous global symmetry that rotates phases of all the fermions relative to all the bosons. This is most easily achieved by a rotation of Grassmann numbers.

$$\theta \rightarrow e^{-i\epsilon} \theta \quad (86)$$

At the same time, one can assign an R-charge for chiral scalar superfield  $\Phi$ :  $R_\Phi$ .

$$\Phi(\theta) \rightarrow e^{i\epsilon R_\Phi} \Phi(e^{-i\epsilon} \theta), \quad A \rightarrow e^{i\epsilon R_\Phi} A, \quad (87)$$

$$\psi \rightarrow e^{i\epsilon(R_\Phi - 1)} \psi, \quad R(\psi) = R(\Phi) - 1 \quad (88)$$

On the other hand, there is no room to rotate the vector superfield, since a nontrivial charge assignment for vector superfield contradicts the nonlinear coupling of vector multiplet in gauge interactions as given in eq.(60). The vector multiplet gives a relative phase rotation between boson and fermion as

$$V(\theta) \rightarrow V(e^{-i\epsilon} \theta), \quad (89)$$

$$\lambda_\alpha \rightarrow e^{i\epsilon} \lambda_\alpha, \quad R(\lambda) = +1 \quad (90)$$

We observe the following characteristic features in the R-symmetry.

1. R-symmetry is chiral. Therefore R-symmetry is generally anomalous.

If there is another anomalous chiral symmetry, usually a linear combination is anomaly free.

2. The mass term for the gaugino  $\lambda$  breaks the R-symmetry

$$\mathcal{L} = \frac{1}{2} m \lambda^\alpha \lambda_\alpha + h.c. \quad (91)$$

3. Superpotential  $P$  must have the R-charge  $R(P) = 2$

$$\mathcal{L} = \frac{1}{2} D^2 P(\Phi) \quad D^2 \approx \frac{\partial^2}{\partial \theta^2} \rightarrow e^{2i\epsilon} D^2, \quad (92)$$

Therefore possible terms in superpotential are restricted if one wishes to have the supersymmetric theory to be invariant under the R-symmetry transformation.

4. Phenomenologically it is desirable to break the R-symmetry explicitly. Since the massless gaugino is not observed in nature, R-symmetry should be broken as is seen from eq.(91). The explicit breaking of the R-symmetry will allow massive gauginos without encountering (light) R-axion resulting from the spontaneous breaking of the R-symmetry. To avoid a rapid proton decay, the R-parity  $(-1)^R$  conservation is desirable replacing the continuous R-symmetry.

### 3 Supersymmetric $SU(3) \times SU(2) \times U(1)$ Model

#### 3.1 Yukawa Coupling

##### 3.1.1 Nonsupersymmetric Standard Model

Let us summarize the nonsupersymmetric  $SU(2) \times U(1)$  model emphasizing the structure of the Yukawa couplings.

We have the (three) generations of the left-handed quark doublets  $q_j$ , the right-handed  $u$ -type quark singlets  $u_{Ri}$ , and the right-handed  $d$ -type quark singlets  $d_{Ri}$ . We also have the (three) generations of the left-handed lepton doublets  $l_j$ , and the right-handed electrons  $e_{Ri}$ . Here  $i, j, \dots$  indicates the generation indices.

We have complex Higgs doublets. Let us denote

$$\begin{pmatrix} \varphi_u \\ \varphi_d \end{pmatrix} \text{ Higgs to give masses to } \begin{pmatrix} u \\ d \end{pmatrix} \text{ type quark}$$

In terms of these fields, the Yukawa couplings  $f$  can be given by

$$L_{Yukawa} = f_u^{ij} \overline{u_{Ri}} \varphi_u^T \varepsilon q_j + f_d^{ij} \overline{d_{Ri}} \varphi_d^T \varepsilon q_j + f_e^{ij} \overline{e_{Ri}} \varphi_d^T \varepsilon l_j \quad (93)$$

where

$$q_i = \begin{pmatrix} u_i \\ d_i \end{pmatrix}, \quad l_i = \begin{pmatrix} \nu_i \\ e_i \end{pmatrix}, \quad (94)$$

$$\varphi_u = \begin{pmatrix} \varphi_u^+ \\ \varphi_u^0 \end{pmatrix}, \quad \varphi_d = \begin{pmatrix} \varphi_d^0 \\ \varphi_d^- \end{pmatrix} \quad (95)$$

$$\varepsilon = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \quad (96)$$

In the nonsupersymmetric model, nothing prevents choosing the Higgs doublet  $\varphi_u$  and  $\varphi_d$  to be the complex conjugate of each other

$$\varphi_u = \varepsilon \cdot \varphi_d^* \quad (97)$$

This is the choice in the nonsupersymmetric minimal standard model.

##### 3.1.2 Supersymmetric Standard Model

It is important to note that the supersymmetric model requires the Yukawa interaction to be a term in the superpotential. This is an  $F$ -type term. The superfield in the Yukawa interaction should have the same chirality.

Therefore we need two Higgs doublet superfields  $H_u$  and  $H_d$  as separate negative chiral scalar superfields.

$$H_u \neq \varepsilon \cdot H_d^* \quad (98)$$

The supersymmetric Yukawa interaction is given by

$$L_{Yukawa} = -\theta P(\Phi)|_{\theta\theta} + h.c. \quad (99)$$

$$\begin{aligned} P = & f_u^{ij} U_i^c H_u^T \varepsilon Q_j + f_d^{ij} D_i^c H_d^T \varepsilon Q_j + f_e^{ij} E_i^c H_d^T \varepsilon L_j \\ & + \mu H_u^T \varepsilon H_d \end{aligned} \quad (100)$$

where we denoted the negative chiral scalar superfield by capital letters and the charge conjugate of the positive chiral scalar superfield in terms of the upper suffix  $c$ .

Higgsino (chiral fermion associated with the Higgs scalar) introduces anomaly in gauge currents. This anomaly has to be cancelled. Introducing the  $H_u$  and  $H_d$  as separate negative chiral scalar superfield serves to achieve the anomaly cancellation at the same time.

### 3.2 Particle Content

Now we find that we need at least a pair of Higgs doublet superfield, we will list the minimal particle content of the supersymmetric standard model. Our convention for the usual standard model  $U(1)$  charge  $Y$  is

$$Q = I_3 + Y \quad (101)$$

The mixing occurs among the following fields

1. Chargino  $\tilde{\varphi}_{u+}$  and  $\tilde{W}^+$
2. Neutralino  $\tilde{\varphi}_{u0}$ ,  $\tilde{\varphi}_{d0}$ ,  $\tilde{W}^0$ ,  $\tilde{B}$
3. Scalar left-right mixing  $\tilde{q}$  and  $\tilde{u}^c, \tilde{d}^c$  etc.

We obtain the following  $R$ -parity  $(-1)^R$  to be conserved and there is no continuous  $R$ -symmetry.

- ordinary particles have  $(-1)^R = +1$
- Supersymmetry particles which are denoted with  $\tilde{\phantom{x}}$ , have  $(-1)^R = -1$

	$J = 1$	$J = 1/2$	$J = 0$	I	Y	$SU(3)$
Gauge fields						
$G$	$g_m$	$\tilde{g}$				
$W$	$W_m$	$\tilde{W}$				
$B$	$B_m$	$\tilde{B}$				
Higgs field						
$H_u = \begin{pmatrix} H_u^+ \\ H_u^0 \end{pmatrix}$		$\tilde{\varphi}_u$	$\varphi_u$	$\frac{1}{2}$	$\frac{1}{2}$	
$H_d = \begin{pmatrix} H_d^0 \\ H_d^- \end{pmatrix}$		$\tilde{\varphi}_d$	$\varphi_d$	$\frac{1}{2}$	$-\frac{1}{2}$	
Quark field						
$Q_i = \begin{pmatrix} U_i \\ D_i \end{pmatrix}$		$q_i$	$\tilde{q}_i$	$\frac{1}{2}$	$\frac{1}{6}$	3
$U_i^c$		$u_i^c$	$\tilde{u}_i^c$	0	$-\frac{2}{3}$	$3^*$
$D_i^c$		$d_i^c$	$\tilde{d}_i^c$	0	$\frac{1}{3}$	$3^*$
Lepton field						
$L_i = \begin{pmatrix} N_i \\ E_i \end{pmatrix}$		$l_i$	$\tilde{l}_i$	$\frac{1}{2}$	$-\frac{1}{2}$	
$E_i^c$		$e_i^c$	$\tilde{e}_i^c$	0	1	
$(N_i^c)$		$\nu_i^c$	$\tilde{\nu}_i^c$	0	0	)

We have denoted the possible right-handed neutrino superfield as  $N_i$ .

## 4 $N = 1$ Supersymmetry Nonperturbative Dynamics

### 4.1 Holomorphy

#### 4.1.1 $N = 1$ Supersymmetry

The chiral scalar superfield contains the complex scalar field  $A$  as the first component as shown in eq.(35).

$$\Phi = A + \sqrt{2}\theta\psi + \theta^2 F \quad (102)$$

The distinction between negative chiral and positive chiral scalar superfield can be formulated as a distinction between holomorphic and anti-holomorphic fields. The former is associated with the complex variable  $z$ , whereas the latter is associated with the complex conjugate variable  $\bar{z}$ .

Since there are terms restricted to the function of the chiral scalar superfield with only one of the chiralities, we obtain a restriction related to the distinction of holomorphic and anti-holomorphic quantities. The principle to distinguish the chirality is called the holomorphy and gives the following restrictions

1. The superpotential is restricted to be a holomorphic function.
2. The Kähler potential and the Fayet-Iliopoulos  $D$ -term are not controlled by holomorphy.

#### 4.1.2 Complexified Symmetry Group

The principle of holomorphy gives the following consequences.

1. If a Lagrangian is invariant under a symmetry group  $G$ , it is automatically invariant under the complexification  $G^c$  of the symmetry group in the case of supersymmetric gauge theories, because of the holomorphy principle.
2. To maintain the supersymmetry, the auxiliary fields have to vanish.

$$F = 0 \quad (103)$$

This is a supersymmetric vacuum condition. One often finds parameters to specify the supersymmetric vacua. These parameters are called moduli.

It has been shown that the moduli in supersymmetric gauge theories are given by gauge invariant holomorphic functions constrained by  $F = 0$  [25].

Because of holomorphy the manifold of vacuum states ( moduli space ) is invariant under complexified symmetry group  $G^c$

3. It is usually most convenient to use the Wess-Zumino gauge to make the physical particle content manifest. The supersymmetric vacuum configuration in the Wess-Zumino gauge is given by the condition that both auxiliary fields should vanish:  $F = 0$  and  $D = 0$ . Since the superpotential is invariant under the complexified symmetry group  $G^c$ ,  $F = 0$  condition is invariant under  $G^c$ . On the other hand, the kinetic term in the Wess-Zumino gauge is invariant under  $G$ , but not invariant under  $G^c$ . Therefore the condition  $D = 0$  is not invariant under  $G^c$ .
4. For NonAbelian gauge group, or Abelian gauge group without the Fayet-Iliopoulos  $D$ -term, it is sufficient to impose the condition  $F = 0$ . Even if the condition  $D = 0$  is not met by the field configuration in  $G^c$ , one can make a complexified gauge transformation to deform  $D$  to vanishing values  $D = 0$ . In this process, the condition  $F = 0$  is unchanged because of the invariance of superpotential under the complexified gauge transformations.

#### 4.1.3 Wilsonian action

In discussing the effective action for low energy field theories, we run across two different kind of the effective potentials.

1. Wilsonian effective action

$$Z = \int D\phi e^{-S_{bare}(\phi, \Lambda)} = \int D\phi_{<} e^{-S_{eff}(\phi_{<})} \quad (104)$$

$$e^{-S_{eff}} \equiv \int D\phi_{>} e^{-S_{bare}(\phi, \Lambda)} \quad (105)$$

We have denoted the modes with momenta larger than the scale  $\mu$  as  $\phi_{>}$ , and the modes with momenta smaller than the scale  $\mu$  as  $\phi_{<}$ .

In this definition, one integrates modes in momentum scales larger than the scale  $\mu$  that one is interested in :  $\phi_{>}$  in  $\mu < p < \Lambda$ . In this definition, one usually suppose that there is a cut-off in the momentum integration to make the integral meaningful and is denoted as  $\Lambda$ . Therefore this can be defined for nonrenormalizable theories as well. This definition has the advantage of receiving no infrared divergences. This feature avoids anomalies to holomorphy. Therefore the Wilsonian effective action  $S_{eff}$  is a holomorphic function of parameters and background fields. It is also noted that the beta function in the Wilsonian action is 1-loop exact in the  $N = 1$  supersymmetric theories [27]. This can most easily be found that the trace anomaly is in the same supermultiplet as the axial anomaly, since the energy-momentum tensor, supercurrent, and the axial current are in the same supermultiplet :

$$(T^{mn}, S_\alpha^m, J^{5m}) \quad (106)$$

On the other hand, the axial anomaly is 1-loop exact according to the Adler-Bardeen theorem [26], whereas the trace anomaly gives the beta function. Therefore the trace anomaly is also one-loop exact provided one does not have anomaly in holomorphy.

## 2. One-Particle-Irreducible (1PI) effective action.

This is the usual effective action in the sense of the generating function for the one particle irreducible amputated amplitudes.

$$Z[J] = \int D\phi e^{-S(\phi) - J\phi} = e^{-W[J]} \quad (107)$$

$$\Phi \equiv \frac{\partial W}{\partial J} \quad (108)$$

$$\Gamma[\Phi] \equiv W[J] - J\Phi \quad (109)$$

If there are massless particles, this effective action usually has an infrared divergences which produces an anomaly for holomorphy. Therefore the beta function in the one particle irreducible effective action receives contributions from all orders of perturbation. More specifically, it can be computed from the knowledge of the one-loop beta function together with the anomalous dimension coming from the wave function renormalization.

$$\beta(\alpha) = -\frac{\alpha^2}{2\pi} \frac{3T(G) - \sum_i T(R_i)(1 - \gamma_i)}{1 - \frac{T(G)\alpha}{2\pi}} \quad (110)$$

$$\gamma_i(\alpha) = -\frac{d \log Z(\mu)}{d \log \mu} = -C_2(R_i) \frac{\alpha}{\pi} + \dots \quad (111)$$

$$T^a T^a = C_2(R) \quad (112)$$

$$\text{tr}(T^a T^a) = T(R) \delta^{ab} \quad (113)$$

## 4.2 Nonperturbative Superpotential

The holomorphy and symmetry requirements restrict the superpotential severely in the case of  $N = 1$  supersymmetric field theories. Quite often these requirements are enough to fix the superpotential  $P$  completely.

On the other hand, the Kähler potential is not holomorphic and is not constrained in the case of  $N = 1$  supersymmetry. Therefore the kinetic term cannot be determined in the  $N = 1$  supersymmetric theories. If we use the  $N = 2$  supersymmetry, however, the kinetic term of the chiral scalar field associated with the vector multiplet is related to the kinetic term of the vector multiplet. Therefore there is a possibility to determine the Kähler potential nonperturbatively.

To find out the results on the nonperturbative effects, let us take the  $SU(N_c)$  gauge group as an example. As for the matter multiplets, we take  $N_f$  flavors of "quark" and "antiquark" chiral scalar superfields  $Q$  and  $\bar{Q}$  in the fundamental representation of  $SU(N_c)$  gauge group.

$$Q_a^i, \quad \bar{Q}_i^a \quad a = 1, \dots, N_c; \quad i = 1, \dots, N_f \quad (114)$$

#### 4.2.1 $N_f < N_c$

Let us consider the massless supersymmetric QCD (SQCD) without superpotential.

$$\mathcal{L}_0 = \int d^4\theta \operatorname{tr}\{Q^\dagger e^{2gV} Q + \tilde{Q} e^{-2gV} \tilde{Q}^\dagger\} \quad (115)$$

$$+ \frac{1}{2} \int d^2\theta \operatorname{tr} W^\alpha W_\alpha + \frac{1}{2} \int d^2\bar{\theta} \operatorname{tr} \bar{W}_{\dot{\alpha}} \bar{W}^{\dot{\alpha}} \quad (116)$$

The global symmetry in this theory at the classical level is given by

$$G_f = SU(N_f)_Q \times SU(N_f)_{\tilde{Q}} \times U(1)_B \times U(1)_A \times U(1)_X \quad (117)$$

Among them there are a number of Abelian global symmetries

$$Q(\theta) \rightarrow e^{i\alpha_B + i\alpha_A} Q(e^{-i\alpha_X} \theta) \quad (118)$$

$$\tilde{Q}(\theta) \rightarrow e^{-i\alpha_B + i\alpha_A} \tilde{Q}(e^{-i\alpha_X} \theta) \quad (119)$$

$$V(\theta) \rightarrow V(e^{-i\alpha_X} \theta) \quad (120)$$

The symmetry  $U(1)_X$  is an  $R$ -type symmetry which make the relative rotation between bosons and fermions.

Let us illustrate how to determine the superpotential.

1. There is an anomaly in  $U(1)_A$  and  $U(1)_X$ .

$$\partial_\mu j^\mu = \frac{1}{32\pi^2} \left[ \sum_i q_i T(R_i) \right] F^a_{\mu\nu} \tilde{F}^{a\mu\nu} \quad (121)$$

$$\operatorname{tr}(t^a t^b) = \frac{1}{2} T(R) \delta^{ab} \quad (122)$$

We can define an anomaly free  $R$ -type symmetry  $U(1)_R$  as a linear combination of  $U(1)_A$  and  $U(1)_X$ . Then the anomaly free  $U(1)$  quantum numbers are listed in the table.

Chiral Field	$U(1)_B$	$U(1)_R$
$Q$	1	$1 - N_c/N_f$
$\tilde{Q}$	-1	$1 - N_c/N_f$

2. Let us next find out the transformation property of the parameter which describes the strength of the gauge interaction  $\Lambda$ .

In order to see this, let us note that there is an instanton solution  $A_{\text{inst}}$

$$F^a_{\mu\nu}(A_{\text{inst}}) = \tilde{F}^a_{\mu\nu}(A_{\text{inst}}) \equiv \frac{1}{2} \varepsilon_{\mu\nu\rho\sigma} F^{\rho\sigma a}(A_{\text{inst}}) \quad (123)$$

In this background, one finds that there are zero modes  $\psi^i_0$  associated with the fermion field  $\psi(x)$  whose number is determined by the index theorem.

$$\gamma^\mu D_\mu(A_{\text{inst}}) \psi_0 = 0 \quad (124)$$

The number of zero modes for a chiral scalar field in the representation  $R$  is  $T(R)$ , which is the second Casimir for the representation. Similarly, the gauge fermion  $\lambda$  has  $T(\text{adj})$  of zero

modes. The effective interaction among fermions can be found by considering the expectation value of an operator  $\mathcal{O}$

$$\begin{aligned}
\langle \mathcal{O} \rangle &= \int D A D \psi D \lambda e^{-S[A, \psi, \lambda]} \mathcal{O} \\
&\approx e^{-S[A_{\text{inst}}]} \int D \psi D \lambda e^{-\psi \gamma^\mu D_\mu(A_{\text{inst}}) \psi - \lambda \gamma^\mu D_\mu(A_{\text{inst}}) \lambda} \mathcal{O} \\
&\approx e^{-S[A_{\text{inst}}]} \int (D \psi D \lambda)_{\text{nonzero}} e^{-(\psi \gamma^\mu D_\mu(A_{\text{inst}}) \psi)_{\text{nonzero}} - (\lambda \gamma^\mu D_\mu(A_{\text{inst}}) \lambda)_{\text{nonzero}}} \\
&\quad \times \prod \int D \psi^i_0 D \lambda^i_0 \mathcal{O}
\end{aligned} \tag{125}$$

where the value of the action at the instanton configuration is given by

$$S[A_{\text{inst}}] = -\frac{8\pi^2}{g^2} \tag{126}$$

Therefore we need to insert appropriate number of fermions in order to have nonvanishing contributions.

$$\left\langle \prod^{T(R)} \psi \prod^{T(\text{adj})} \lambda \right\rangle \propto \exp\left(-\frac{8\pi^2}{g^2} + i\theta\right) = \Lambda^{3N_c - N_f} \tag{127}$$

where the coefficient of the one-loop beta function is given by  $b = 3N_c - N_f$ .

3.  $U(1)_A$  transformation property of fermions are given for quarks and antiquarks by

$$\psi \rightarrow e^{i\alpha q} \psi, \quad q = 1 \tag{128}$$

for gauginos

$$\lambda \rightarrow e^{i\alpha q_\lambda} \lambda, \quad q_\lambda = 0 \tag{129}$$

Therefore the theory can be made invariant provided we assign the transformation property for the parameter  $\Lambda$  as

$$\left\langle \prod \psi \prod \lambda \right\rangle \rightarrow e^{i\alpha(2N_f q T(R) + q_\lambda T(\text{adj}))} \left\langle \prod \psi \prod \lambda \right\rangle \tag{130}$$

The above result shows that the theory itself is not invariant under this  $U(1)_A$  transformation. Therefore it is anomalous. The amount of the anomaly is such that we can relate the (different) theory by assigning the above transformation property to the parameter of the theory,  $\Lambda$ . By this transformation, we are relating different theories. This property becomes useful when we determine the nonperturbative superpotential.

Therefore if we transform the parameter of the theory  $\Lambda$  as if it is a background field, we arrive at another theory related by the symmetry transformation. Hence there is a family of theories that are related by the transformations and the predictions of the theories are related by the transformation.

$$\Lambda^{3N_c - N_f} \rightarrow e^{i\alpha(2N_f q T(R) + q_\lambda T(\text{adj}))} \Lambda^{3N_c - N_f} \tag{131}$$

Namely  $\Lambda^{3N_c - N_f}$  can be regarded as having  $U(1)_A$  charge  $2N_f q T(R) + q_\lambda T(\text{adj}) = 2N_f$ .

One should imagine that the parameter to be a kind of background fields when one considers the transformation of the parameter of the theory. This method has been used extensively by Seiberg and collaborators [28].

4. Let us constrain the superpotential of the low energy effective action by demanding several requirements successively. The principle of holomorphy requires that the superpotential has to be a function of negative chiral scalar superfields only. Gauge invariance requires that



the superpotential should be a function of gauge invariant combinations of superfields. Since  $M_i^j = \tilde{Q}^a_i Q_a^j$  is the only color singlet negative chiral scalar superfield for the case  $N_c > N_f$ , we find that the superpotential should be a function of  $M_i^j = \tilde{Q}^a_i Q_a^j$ . Let us note that the holomorphy forbids to use the gauge invariant combination of negative and positive chiral scalar superfields such as  $(Q_a^i)^* Q_a^j$ . The global symmetry  $SU(N_f) \times SU(N_f)$  dictates that the effective superpotential  $P$  should be a function of  $\det(Q\tilde{Q})$  only.

$$P(Q, \tilde{Q}) = f\left(\det(Q\tilde{Q})\right) \quad (132)$$

Next we can use the transformation property under the (anomalous) global  $U(1)_A$ . As we have seen, the effective superpotential should be invariant under the transformation provided we assign a  $U(1)_A$  charge for the parameter  $\Lambda^{3N_c-N_f}$  as  $2N_f q T(R) + q_\lambda T(adj) = 2N_f$ . Therefore the superpotential should contain the parameter  $\Lambda$  as a function of the ratio  $\Lambda^{3N_c-N_f} / \det(Q\tilde{Q})$  only. The dimensional analysis gives that the superpotential has to have the dimensions of  $M^3$ . Thus superpotential is determined except overall numerical constants  $C_{N_c N_f}$  that depend on  $N_c$  and  $N_f$ .

$$P = C_{N_c N_f} \left[ \frac{\Lambda^{3N_c-N_f}}{\det(Q\tilde{Q})} \right]^{\frac{1}{N_c-N_f}}, \quad (133)$$

This set of numerical constants can be determined by two consistency conditions regarding the decoupling:

- (a) If we give a large mass to a quark  $Q_{N_f}$ , it should decouple. This relates the  $N_f$  case with  $N_f - 1$  case with  $N_c$  unchanged.
- (b) If we give a large vacuum expectation value to a squark  $Q_i$ , the color gauge symmetry is partially broken and part of the flavor is decoupled. This relates the  $N_c, N_f$  case with  $N_c - 1, N_f - 1$  case.

These two consistency conditions reduce the numerical coefficients to a single number  $C$ .

$$C_{N_c N_f} = (N_c - N_f) C^{\frac{1}{N_c-N_f}} \quad (134)$$

We can see that the  $\Lambda$  dependence of the  $N_f = N_c - 1$  case agrees exactly with the one instanton contribution. Since the gauge symmetry is broken completely in this case, we can consider the large vacuum expectation values which corresponds to the weak coupling situation. Therefore we can trust the one-instanton calculation in this case and find

$$C = 1 \quad (135)$$

The resulting nonperturbative exact superpotential can be summarized as

$$P_{np} = \epsilon_{N_c-N_f} (N_c - N_f) \left[ \frac{\Lambda^{3N_c-N_f}}{\det(Q\tilde{Q})} \right]^{\frac{1}{N_c-N_f}} \quad (136)$$

$$(\epsilon_{N_c-N_f})^{N_c-N_f} = 1 \quad (137)$$

If we consider the large vacuum expectation values for all the quark flavors, the gauge symmetry is broken from  $SU(N_c)$  to  $SU(N_c - N_f)$ . The effective coupling between these two gauge theories should match at the scale of the vacuum expectation values. This matching condition reads

$$\left( \frac{\Lambda_{N_c, N_f}}{(\det \tilde{Q} Q)^{\frac{1}{2N_f}}} \right)^{3N_c-N_f} = \left( \frac{\Lambda_{N_c-N_f, 0}}{(\det \tilde{Q} Q)^{\frac{1}{2N_f}}} \right)^{3(N_c-N_f)} \quad (138)$$

For  $N_f \leq N_c - 2$ ,

$$-\frac{8\pi^2}{g^2(\mu)} = \log \left( \frac{\Lambda}{\mu} \right)^b, \quad b = 3N_c - N_f \quad (139)$$

$$\begin{aligned} \mathcal{L} &= \frac{1}{4g^2} \int d^2\theta W^\alpha W_\alpha + \dots \\ &= -\frac{1}{32\pi^2} \log \left( \frac{\Lambda}{\mu} \right)^b \int d^2\theta W^\alpha W_\alpha + \dots \end{aligned}$$

The first component of the superpotential corresponds to the gaugino bilinear. Therefore the nonperturbative superpotential can be understood as gaugino condensation in the unbroken gauge group  $SU(N_c - N_f)$

$$\frac{1}{32\pi^2} \langle 0 | \lambda^\alpha \lambda_\alpha | 0 \rangle = \epsilon_{N_c - N_f} \Lambda_{N_c - N_f, 0}^3 \quad (140)$$

So far we have discussed the nonperturbative effects in the  $N = 1$  supersymmetric gauge theories. There has been much progress in recent years on the nonperturbative effects not only for the  $N = 1$  supersymmetric theories but also for higher  $N$  supersymmetric theories that we have not enough space to cover. Among them it is worth mentioning that the exact solution for the low energy effective action of  $N = 2$  supersymmetric gauge theories has been obtained up to two derivatives including the full nonperturbative effects [29].

## 5 Summary

1. Supersymmetry is the most promising solution to the gauge hierarchy problem.
2. Supersymmetry is the only nontrivial relativistic symmetry that relates particles with different spin.
3. Good progress has been made to understand the nonperturbative dynamics of supersymmetric gauge theories in both  $N = 1$  and  $N = 2$  supersymmetric theories.

## Appendix A. Spinors and conventions

Our convention for the metric is given by  $\eta_{mn} = (-1, +1, +1, +1)$ . The  $\gamma$  matrices are defined in our convention by ( $\gamma_m^{\text{here}} = \gamma_m^{\text{Wess-Bagger}} = \gamma_m^{\text{Bjorken-Drell}}$ )

$$\gamma_m \gamma_n + \gamma_n \gamma_m = -2\eta_{mn} \quad (1)$$

The conjugate spinor  $\bar{\psi}$  for the spinor  $\psi$  is given by  $\bar{\psi} \equiv \psi^\dagger \gamma_0 = -\psi^\dagger \gamma^0$ . The chiral  $\gamma$  matrix  $\gamma_5$  is defined by

$$\gamma_5 = \gamma^5 = \gamma^0 \gamma^1 \gamma^2 \gamma^3 = \gamma_5^{\text{Wess-Bagger}} = -i\gamma_5^{\text{Bjorken-Drell}} \quad (2)$$

It is useful to use the Weyl basis of  $\gamma$  matrix

$$\gamma_0 = -\gamma^0 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \gamma_j = \gamma^j = \begin{pmatrix} 0 & \sigma^j \\ -\sigma^j & 0 \end{pmatrix}, \quad j = 1, 2, 3 \quad (3)$$

Combined together we introduce four dimensional notation for the two by two matrices  $\sigma^m, \bar{\sigma}^m$

$$\gamma^m = \begin{pmatrix} 0 & \sigma^m \\ \bar{\sigma}^m & 0 \end{pmatrix}, \quad \sigma^0 = \bar{\sigma}^0 \equiv -1, \quad \bar{\sigma}^j = -\sigma^j \quad (4)$$

In this basis, the chiral  $\gamma$  matrix becomes diagonal

$$\gamma_5 = -i \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \quad (5)$$

Since supersymmetry is conveniently formulated in terms of spinors of definite chirality, it is useful to decompose the usual four component spinor into upper and lower two component spinors with the definite chirality.

$$\psi \equiv \begin{pmatrix} \xi_\alpha \\ \eta^{\dot{\alpha}} \end{pmatrix} \equiv \begin{pmatrix} \xi_\alpha \\ \bar{\eta}^{\dot{\alpha}} \end{pmatrix} \quad (6)$$

The negative and positive chirality spinors have undotted and dotted indices which are raised and/or lowered by antisymmetric  $\epsilon$  tensor

$$\epsilon^{12} = -\epsilon_{12} = 1, \quad \epsilon_{\alpha\beta}\epsilon^{\beta\gamma} = \delta_\alpha^\gamma \quad (7)$$

The conjugate spinor is given by

$$\bar{\psi} = ( (\bar{\eta}^{\dot{\alpha}})^* \quad (\xi_\alpha)^* ) = ( \eta^\alpha \quad \xi^*_{\dot{\alpha}} ) \equiv ( \eta^\alpha \quad \bar{\xi}_{\dot{\alpha}} ) \quad (8)$$

$$\xi^\alpha \equiv \epsilon^{\alpha\beta} \xi_\beta, \quad \eta_{\dot{\alpha}} \equiv \epsilon_{\dot{\alpha}\dot{\beta}} \eta^{\dot{\beta}} \quad (9)$$

The charge conjugation matrix  $C$  is defined by

$$C^{-1} \gamma^m C = -\gamma^{mT} \quad (10)$$

One can show that  $C$  is antisymmetric and can be chosen to be unitary  $C^T = -C$ ,  $C^\dagger C = 1$ . In the two-component notation using the Weyl basis, we have

$$C = -i\gamma_2\gamma_0 = \begin{pmatrix} -i\sigma^2 & 0 \\ 0 & i\sigma^2 \end{pmatrix} = \begin{pmatrix} \epsilon_{\alpha\beta} & 0 \\ 0 & \epsilon^{\dot{\alpha}\dot{\beta}} \end{pmatrix} \quad (11)$$

The charge conjugate spinor corresponds to antiparticle and is defined by

$$\psi^c \equiv C\bar{\psi}^T, \quad \bar{\psi}^c = -\psi^T C^{-1} \quad (12)$$

The charge conjugation reverses the chirality

$$\psi = \begin{pmatrix} \xi_\alpha \\ \bar{\eta}^{\dot{\alpha}} \end{pmatrix} \rightarrow \psi^c = \begin{pmatrix} \epsilon_{\alpha\beta} \eta^\beta \\ \epsilon^{\dot{\alpha}\dot{\beta}} \bar{\xi}_{\dot{\beta}} \end{pmatrix} = \begin{pmatrix} \eta_\alpha \\ \bar{\xi}^{\dot{\alpha}} \end{pmatrix} \quad (13)$$

Spinors which are charge conjugate of itself is called the majorana spinor

$$\psi^c = \psi \rightarrow \psi = \begin{pmatrix} \eta_\alpha \\ \bar{\eta}^{\dot{\alpha}} \end{pmatrix} \quad \bar{\psi} = ( \eta^\alpha \quad \bar{\eta}_{\dot{\alpha}} ) \quad (14)$$

## Appendix B. Grassmann number and its derivatives

Grassmann number is defined as the anticommuting c-number. The derivative in terms of Grassmann number is defined by

$$\frac{\partial}{\partial \psi_\alpha} \psi_\beta = \delta_{\alpha\beta}, \quad \frac{\partial}{\partial \bar{\psi}_\alpha} \bar{\psi}_\beta = \delta_{\alpha\beta} \quad (1)$$

$$\frac{\partial}{\partial \psi_\alpha} \bar{\psi}_\beta = (C^{-1})_{\beta\alpha}, \quad \frac{\partial}{\partial \bar{\psi}_\alpha} \psi_\beta = (C)_{\beta\alpha} \quad (2)$$

$$\frac{\partial}{\partial \psi_\alpha} = \frac{\partial}{\partial \bar{\psi}_\beta} (C^{-1})_{\beta\alpha}, \quad \frac{\partial}{\partial \bar{\psi}_\alpha} = -(C)_{\alpha\beta} \frac{\partial}{\partial \psi_\beta} \quad (3)$$

$$\bar{\epsilon} \frac{\partial}{\partial \theta} = - \frac{\partial}{\partial \bar{\theta}} \epsilon \quad (4)$$

Two-component notation

$$\frac{\partial}{\partial \eta_\alpha} \eta_\beta = \delta_\beta^\alpha, \quad \frac{\partial}{\partial \bar{\eta}_{\dot{\alpha}}} \bar{\eta}^{\dot{\beta}} = \delta_{\dot{\alpha}}^{\dot{\beta}} \quad (5)$$

$$\frac{\partial}{\partial \eta_\alpha} \eta^\beta = \epsilon^{\beta\alpha}, \quad \frac{\partial}{\partial \bar{\eta}_{\dot{\alpha}}} \bar{\eta}_{\dot{\beta}} = \epsilon_{\dot{\beta}\dot{\alpha}} \quad (6)$$

$$\frac{\partial}{\partial \eta^\alpha} \eta_\beta = \epsilon_{\beta\alpha}, \quad \frac{\partial}{\partial \bar{\eta}_{\dot{\alpha}}} \bar{\eta}^{\dot{\beta}} = \epsilon^{\dot{\beta}\dot{\alpha}} \quad (7)$$

$$\frac{\partial}{\partial \eta_\alpha} = \frac{\partial}{\partial \eta^\beta} \epsilon^{\beta\alpha}, \quad \frac{\partial}{\partial \bar{\eta}_{\dot{\alpha}}} = \frac{\partial}{\partial \bar{\eta}_{\dot{\beta}}} \epsilon^{\dot{\beta}\dot{\alpha}} \quad (8)$$

$$\frac{\partial}{\partial \eta^\alpha} = -\epsilon_{\alpha\beta} \frac{\partial}{\partial \eta_\beta}, \quad \frac{\partial}{\partial \bar{\eta}_{\dot{\alpha}}} = -\epsilon^{\dot{\alpha}\dot{\beta}} \frac{\partial}{\partial \bar{\eta}_{\dot{\beta}}} \quad (9)$$

$$\epsilon^\alpha \frac{\partial}{\partial \theta^\alpha} = - \frac{\partial}{\partial \theta_\alpha} \epsilon_\alpha, \quad \bar{\epsilon}_{\dot{\alpha}} \frac{\partial}{\partial \bar{\theta}_{\dot{\alpha}}} = - \frac{\partial}{\partial \bar{\theta}_{\dot{\alpha}}} \bar{\epsilon}^{\dot{\alpha}} \quad (10)$$

## References

- [1] H. Georgi and H. Glashow, *Phys. Rev. Lett.* **32** (1974) 438.
- [2] H. Georgi, H. Quinn and S. Weinberg, *Phys. Rev. Lett.* **33** (1974) 451.
- [3] M. Veltman, *Acta Phys. Pol.* **B12** (1981) 437.
- [4] G. 't Hooft, in Recent Developments in Gauge Theories, Cargèse summer school 1979 p.135.
- [5] L. Susskind, *Phys. Rev.* **D20** (1979) 2619; S. Weinberg, *Phys. Rev.* **D19** (1979) 1277; **D13** (1976) 974; S. Dimopoulos, and L. Susskind, *Nucl. Phys.* **B155** (1979) 237; E. Eichten and K. Lane, *Phys. Lett.* **B90** (1980) 125.
- [6] J. Wess and J. Bagger, *Supersymmetry and Supergravity*, Princeton University Press, (1992).
- [7] N. Sakai, *Z. f. Phys.* **C11** (1981) 153.
- [8] S. Dimopoulos and H. Georgi, *Nucl. Phys.* **B193** (1981) 150.
- [9] E. Witten, *Nucl. Phys.* **B188** (1981) 513.
- [10] R. Kaul, *Phys. Lett.* **B109** (1982) 19.
- [11] P. Fayet, *Phys. Lett.* **B69** (1977) 489.
- [12] For a review on supersymmetric models, see for instance, H.P. Nilles, *Phys. Rep.* **C110** (1984) 1; P. Nath, R. Arnowitt, and A. Chamseddine, *Applied N = 1 Supergravity*, the ICTP Series in Theoretical Physics, Vol.I (World scientific) 1984; H. Haber and G. Kane, *Phys. Rep.* **C117** (1985) 75.
- [13] For a selection of papers on supersymmetric theories including models, see for instance, S. Ferrara, *Supersymmetry*, Vol.I and II (World scientific) 1987.
- [14] U. Amaldi, W. de Boer, and H. Fürstenau, *Phys. Lett.* **B260** (1991) 447.
- [15] N. Arkani-Hamed, S. Dimopoulos, and G. Dvali, *Phys. Lett.* **B429** (1998) 263.

- [16] G. Parisi and N. Surlas, *Phys. Rev. Lett.***43** (1979) 744.
- [17] S. Coleman and J. Mandula, *Phys. Rev.* **159** (1967) 1251.
- [18] R. Haag, J. Lopuszanski, and M. Sohnius, *Nucl. Phys.* **B88** (1975) 257.
- [19] E. Wigner, *Ann. Phys.***40** (1939) 149.
- [20] M.K. Prasad and C.M. Sommerfield, *Phys. Rev. Lett.***35** (1975) 760; E.B. Bogomolny, *Sov. J. Nucl. Phys.* **24** (1976) 449.
- [21] E. Witten and D. Olive, *Phys. Lett.* **78B** (1978) 97.
- [22] M.T. Grisaru, W. Siegel, and M. Rocek, *Nucl. Phys.***B159** (1979) 420.
- [23] K. Fujikawa and W. Lang, *Nucl. Phys.***B88** (1975) 61.
- [24] P.S. Howe, K.S. Stelle, and P.C. West, *Phys. Lett.* **124B** (1983) 55.
- [25] M.A. Luty and Washington Taylor IV, *Phys. Rev.* **D53**(1996) 3399.
- [26] S.L. Adler and W.A. Bardeen, *Phys. Rev.* **182**(1969) 1517.
- [27] M.A. Shifman and A.I. Vainshtein, *Nucl. Phys.* **B277** (1986) 456, **B359** (1991) 571.
- [28] A summary on the the  $N = 1$  supersymmetric theories has been gievn in K. Intriligator and N. Seiberg, Lectures on supersymmetric gauge theories and electric-magnetic duality, hep-th/9509066.
- [29] N. Seiberg and E. Witten, *Nucl. Phys.* **B426** (1994) 19, **B431** (1994) 484.

# 22. Conformal Field Theory: A Bridge Over Troubled Waters

W. Nahm \*

Physikalische Institut, Bonn University

## Abstract

A perspective overview of QFT from its Dirac beginning to modern string theory is sketched, with special reference to the role of Conformal Field Theory in bridging the uneasy relationship between mathematics and physics through this entire century.

## 1 Introduction

In his 1972 address to the American Mathematical Society, Dyson deplored the 'divorce' between mathematics and physics over the issue of quantum field theory. The present book on the impact of field theory on modern physics, timed in accordance with the International Mathematics Year of 2000 AD, gives hope that the rift will soon be bridged. In conformal quantum field theory in two dimensional spacetime, conditions are particularly favorable for gaining common ground. This area can attract both mathematicians and physicists by its beauty, its transparent mathematical structure and its many applications. Thus its investigation has moved to a rather central position in quantum field theory as a whole. Perhaps one can say that a bridge exists since many years, but has been used much below its capacity.

Analogues in more than two dimensions will be mentioned in the present article, but they have not been developed very far nor found mathematical applications yet. For convenience, the expression 'conformal field theory' will refer to conformally invariant quantum field theories in two dimensions, if nothing else is said.

Its applications are surprisingly diverse. In mathematics, one particular theory has become well known due to its automorphism group, the Fischer-Griess monster, and by the Fields medal for Borcherds. Currently research on the many mirror symmetric cases makes rapid progress, whereas the explicit construction of Kähler-Einstein metrics is just at a planning stage. In physics, conformal field theory became essential for the study of continuous phase transitions in condensed matter physics, but the most important applications concern string theory. Even independently of its status as the best candidate for a theory of quantum gravity, string theory has become an important tool in quantum field theory. In particular, it yields a map from two-dimensional conformal theories to quantum field theories in higher dimensions, even to rather realistic examples in four-dimensional spacetime.

It will be argued that conformal field theory can satisfy all mathematicians who want to understand quantum field theory without giving up their standards of clarity and rigour. In return, many of its important aspects need advanced mathematical techniques. Some still remain out of the reach of physicists, others are handled in a manner which produces lots of undiscovered errors in the literature. A serious involvement of mathematicians will yield much firmer foundations for the things to come.

Apart from conformal field theory, this article will discuss some string theory. The perturbative aspects of the latter are well understood by now and should be easily accessible. Indeed, their bulk

---

\*Email: werner@avzw01.physik.uni-bonn.de

just constitutes a direct application of conformal field theory. By themselves, these aspects do not yet transcend the old principles of 20th century physics, but they are incomplete and seem to point in one unique direction of progress. There, non-perturbative string theory (or M-theory, or however one chooses to call it) has started to emerge and should lead to a new insight. Its birth might be a lot easier if mathematicians again get into the habit of acting as good midwives.

There is no lack of good will. Many conferences have been attended by mixed groups of mathematicians and physicists. Most importantly, Princeton has brought together many of the best people of both communities in a dedicated program. Nevertheless, one sometimes gets the feeling that a new generation will be needed to overcome the difficulties.

It may help to recognize what the main obstacles have been, since a discussion of past successes and mistakes can prepare the way for future achievements. In any case, the present article is supposed to cover the early stages of the discoveries. Of course, history is complex, full of turbulence and countercurrents, such that precise statements would need more hedging or much more research. The historical content of the present article should therefore be taken as a signpost, not as a map. Its indications will be approximate, but may be helpful. Moreover, a low density of formulas should make the article accessible to a wider readership. Methodological qualms of the historians will be brazenly ignored. If they deny that it makes sense to ask what might have happened, what about an advanced quantum computer which reruns history, starting from a reasonable subspace of initial conditions.

An invitation for mathematicians to cross the bridge also needs some technical parts, however. Mathematicians still think of quantum field theory as a useful source of ideas (cf. the Seiberg-Witten equations), but otherwise as impenetrable, though some related structures are regarded as good mathematics (topological quantum field theory, probably conformal field theory). From the point of view of a physicist, this is a strange attitude. By nature, conformal field theory was presented in the context of quantum field theory as a whole, so this is the way it should be discussed. Nature has a habit of posing such problems, recall infinitesimals and differential equations. Again, we should have confidence in her guidance.

The article starts with a short introduction to the history of quantum field theory. For readers who want to get a fuller picture and the necessary references, there are good reviews and reprint volumes at different technical levels [Schwinger 1958, Pais 1986, Crease and Mann 1987]. In those reviews the aim is to show how nature was explained. Here, mathematically well-built structures will be regarded as equally important for further progress. We shall see that the tools to build the bridge were available at the end of the 60's, maybe even ten years earlier.

Because of the focus on conformally invariant theories, other issues of mathematically rigorous quantum field theory, like the work of Glimm and Jaffe on superrenormalizable theories will not be discussed, however. Even for the main themes of the article, the selection is partial. Current algebras are one of the major themes of conformal field theory, but here the collaboration between mathematicians and physicists has been very fruitful since twenty years, and there are already nice reviews, e.g. [Goddard and Olive 1988], so the topic will receive less emphasis than would be necessary in a complete survey. Renormalization will be discussed in some detail, since most mathematicians regard it as the major stumbling block which prevents an understanding of quantum field theory. Thus it may help to see that this procedure is rather easy from a mathematical point of view and follows a well known idea of the 19th century. Path integrals will not be mentioned. In quantum field theory as a whole the corresponding ideas have not yet found a satisfactory form, and even in conformal field theory one needs a discussion in terms of categories and operads, which only seems simple to the very good or the very young. The discussion of string theory is limited to the perspective of conformal field theory. A serious consideration of its non-perturbative aspects would have to start with a description of instantons and solitons. Many connections with conformal invariance could be explained, but not within the scope of the present article. Altogether, it is unavoidable that many readers will have reasons to complain, but at least they should feel encouraged to do better.

## 2 After a golden age

A theoretical physicist who looks back to the beginning of the past century has reasons to feel rather humble. A few decades witnessed three revolutionary insights in structures of nature, namely special relativity, quantum mechanics and Einstein's theory of gravity. We still have to embed these in a unified theory. Meanwhile, we covered much territory in the study of the complexities of nature, but further understanding of her basic features proceeds at a snail's pace.

Evidently, the initial rapid progress relied on a close interaction between physicists and mathematicians. We cannot even talk about those three structures without alluding to this fact. Just think of the Poincaré group, Hilbert space and Riemannian geometry.

In the first years of the century, many physicists felt doubtful or uneasy about the importance of contemporary mathematical methods. For example, Mittag-Leffler struggled in vain to get a Nobel prize for Poincaré out of the old establishment. Around 1920, the movement had become irresistible, however. Einstein's Nobel prize document avoids to mention special or general relativity, but this was hardly more than a funny detail. Everyone went to Göttingen, Weyl solved the Schrödinger equation for the hydrogen atom, and Hilbert was an eager competitor when Einstein approached the final form of his equation for gravity. Of longer lasting importance was the clarification of the mathematical structure of quantum mechanics by v. Neumann and Weyl. Von Neumann had to immerse himself in physics because of the bomb and the computer, but Weyl's publication list in physics journals is impressive, too, and his exchange of ideas with Einstein and Pauli was particularly fruitful.

Mathematicians made contributions of three kinds. Sometimes they solved concrete problems which the physicists found too difficult, but this was rare. More importantly, the internal logic of mathematics had led to the discovery of deep structures which found unexpected applications. Finally, the analysis of discoveries made in physics uncovered new mathematical worlds and allowed the physicists to think more clearly and efficiently about their own results.

It is no surprise that the latter task attracts some of the best mathematical minds. Already in 1900, Hilbert saw that a renewed interest in physics would be productive, and he put the development of good axioms for mechanics as the sixth problem on his famous list. Twenty-five years later, classical mechanics had the necessary clear conceptual basis to allow the wonderful emergence of quantum mechanics. In contrast, the infinite dimensional spaces of classical field theory remained less well understood. This contributed a bit to the confusion about quantum field theory, as we shall see.

When a science is restructured during and after a major advance, it is particularly important to put what has been known before in a new context and to provide it with deeper foundations. Conformal invariance in physics emerged in this way, though at first its place seemed to be marginal. Its discovery was triggered by special relativity. This theory had underlined the importance of symmetry groups and stimulated a new mathematical look at Maxwell's equations. Cunningham [1910] and Bateman [1910] determined the maximal group of symmetries of the latter and discovered their conformal invariance. In other words, Maxwell's system of equations is invariant under the maximal group of spacetime transformations which preserve angles, but not necessarily distances. Addressing mathematics and physics audiences, F. Klein repeatedly asked for an explanation, but in vain.

Some progress became possible with the work of E. Noether [1918]. She was a creator of modern abstract algebra and only had a marginal interest in physics, but was motivated by Einstein's gravity theory to explain the general connection between symmetries of Lagrangians and conservation laws. Prompted by Klein to analyse the problem with Noether's method, Bessel-Hagen determined the conserved quantities corresponding to the conformal invariance of Maxwell's equations [1921]. In the following years, several other differential equations were investigated. Most importantly, Pauli proved that the Dirac equation is conformally invariant when the mass vanishes [1940].



By that time, Hodge had found the tools for a deeper analysis [1941], which needed much longer to make headway into the physics community. He had investigated topological questions like Poincaré duality from the point of view of differential forms. In this language, the electromagnetic field can be described by a 2-form  $F$ . In the absence of charges, Maxwell's equations take the form

$$dF = 0, \quad d * F = 0.$$

This simple form even applies to the curved spacetime of Einstein's theory. Indeed, the differential operator  $d$  does not depend on any kind of metric structure. The Hodge duality operator  $*$ , which acts linearly on differential forms, depends on the metric. When applied to  $k$ -forms in a spacetime of  $2k$  dimensions, however, the dependence on the distance scale cancels out. In particular, in the most important case of four-dimensional spacetime,  $*dF$  only depends on the angles, in other words on the conformal structure. Moreover, the Lagrangian density of the electromagnetic field in empty space is given by the integral over  $F \wedge *F$ . Since the metric again appears only through the Hodge star operation, the Poisson brackets derived from this Lagrangian have the same conformal invariance. This remains true for the corresponding quantum system, but in general conformal invariance is broken when one introduces charges.

Eventually, this approach turned out to be very productive for physics, see e.g. [Atiyah, Hitchin, Singer 1978], but mainstream physicists learned about it only in the 70's, largely through the efforts of Atiyah, who had heard Hodge's lectures as a student. Indeed, in the 30's a fault-line between the communities of the physicists and the mathematicians had started to develop. It must have been hard to spot at the time, since there were greater and more immediate concerns. Within physics, the split between theoreticians and experimentalists now became complete. Einstein still had done moderately respectable experimental work, but Heisenberg's PhD exam was a near disaster, since he had concentrated all his efforts on theory. Some misgivings of the experimentalists were quite natural. Though there was no reason to expect that mathematical physics would regain the prestigious position it had had in the times of Newton, Euler and Lagrange, the incredible popularity of Einstein and the difficulties of his theories must have suggested to many that something had gone wrong.

Finally, from the 30's to the 50's physics was hopelessly entangled in far more important events (fascism, the war, the bomb, Stalin, McCarthy, ...). This was a worldwide phenomenon, which left little refuge. It had one important positive aspect, however. Due to the international contacts, progress in physics was no longer the prerogative of Europe or North America. Most visibly, Japan and India had started to take part.

All of this left little room for concern about the spreading rift between theoretical physics and mathematics. Around 1970, however, it had become too big to overlook. In his Gibbs Lecture, Dyson put it bluntly: "the marriage between mathematics and physics ... has recently ended in divorce" [Dyson 1972].

The main culprit seemed to be quantum field theory. Here is a fairly typical quotation from a highly regarded textbook: "The mathematically inclined reader undoubtedly by now will have had serious misgivings about the validity and meaningfulness of the renormalization program, since this program has at its point of departure a set of meaningless equations which it then proceeds to manipulate according to rules which are outside the bounds of conventional mathematics to obtain (presumably) finite results (not to mention the fact these prescriptions, as outlined in the present chapter, are applicable only to the power series expansion of the 'meaningless equations,' which power series expansion in all probability does not converge!)" [Schweber 1964, p. 645].

It is clear that something had gone wrong. In a sense, one may put the blame on nature, since she gave ambiguous directions. We considered the discoveries of special relativity, quantum mechanics and Einstein's theory of gravity, but it is somewhat misleading to talk about them on a par, since the three theories do not occupy the same logical level. Einstein's name for his theory of gravity was general relativity, because compatibility with the principles of special relativity was incorporated from its inception. Thus the task of unification would be finished, if one could join

gravity and quantum mechanics in one move. From today's point of view, this problem was too difficult and led into a thick fog.

Instead one could follow the geometrical path indicated by Einstein's gravity theory. This was natural for mathematicians, but not immediately productive for physics. Nevertheless, the search in these directions provided a favorable environment for the development of gauge theories, as we shall see later.

For physicists, a different path was indicated by nature. After quantum mechanics had matured around 1926 (the year of the Schrödinger equation), the next fundamental problem was to put together quantum theory and special relativity. The essential guidance came from the experiments, whereas the mathematical structures remained rather obscure. In favorable circumstances, it still might have been possible to advance together, but many of the links between mathematics and physics were broken by the war. When a deeper study of the weak and strong interactions led to gauge theories, a convergence of the two paths was indicated, but this came too late for a reestablishment of the old contacts. In this sense the lack of mathematical accessibility of relativistic quantum field theory is rather a consequence of the separation of mathematics and physics than its cause.

Let us come back to the perspective of 1926. Classical physics deals with rigid bodies and with fields. Now the former were to be regarded as a low velocity approximation, since extended rigid bodies are incompatible with special relativity. When an object is touched, it cannot be affected all at once, since this would surpass the speed of light.

Rigid bodies often had been approximated by point particles. Now they had to be considered in terms of pointlike constituents. In one space dimension, the latter can interact by collisions, but in more dimensions this makes little sense. Thus the only available candidates for the description of interactions in the real world were field theories. Conversely, the discovery of special relativity depended on an analysis of Maxwell's equations for the electromagnetic field. In quantum physics, matter in the form of point particles was easily incorporated in this frame, since the Schrödinger wave function could be regarded as the avatar of a relativistic field. Thus the unification of special relativity and quantum theory demanded the formulation of quantum field theory.

These ideas were well understood in the 1920's. They came very naturally, since already the first steps of quantum mechanics were guided by quantum field theory: The first formula for a quantal phenomenon was Planck's radiation law for the electromagnetic fields emitted by a heated black body. Thus quantum electrodynamics took shape immediately after the birth of modern quantum mechanics, in a 1927 paper by Dirac and, in more appropriate guise, in a paper by Heisenberg and Pauli in 1929. In its initial form, it was sufficient for a calculation of the semiclassical electromagnetic processes observable at that time.

Soon, however, quantum field theory was put in doubt by experimental results and problems of consistency. For experimentalists, further work on the unification of special relativity and quantum theory posed a single basic challenge - study particle interactions at velocities close to the speed of light. Here early researchers were confronted with a bewildering wealth of data, from nuclear physics to cosmic rays. It took a long time until things were sorted out to reveal the underlying structures. In particular, it was far from obvious that the experimental results could be described by any kind of quantum field theory. For a long time, electromagnetism was the only interaction for which it made real sense.

For mathematicians and physicists alike, this greatly diminished the attractiveness of quantum field theory. Most importantly, it contributed to the persistent but unproductive expectation of another revolution in the foundations of physics. In view of the previous decades, this expectation was quite understandable. Quantum mechanics had been developed for atomic physics, particle physics might need something equally revolutionary and exciting. Oppenheimer even gave a number: don't believe the old ideas beyond 100 MeV. For a while, there was a concrete reason for this attitude. Yukawa had predicted a particle of 100 MeV to explain the strong interaction, but when it seemed to be discovered, most of its other properties were wrong. Eventually, muons and

pions were distinguished and the paradoxes dissolved away, but the basic attitude surfaced again on many occasions.

Quantum electrodynamics itself set other obstacles against the joint development of quantum field theory in a common effort of physicists and mathematicians. In particular, one immediately had to face one of the old problems of classical physics, namely the infinite energy in the electric field of point charges. In classical physics, spreading out the charge yielded a temporary excuse, but special relativity and quantum mechanics demanded the consideration of point charges, such that a clash was inevitable. There soon came a reason for hope, however. Against his expectations, Weisskopf (with a little help from Furry) showed in 1934 that in quantum electrodynamics the pole divergence of the classical theory is replaced by a mild logarithmic one. In the following years, Kramers explained the basic principles of regularization and renormalization. Weisskopf apparently was slowed down by discontent about his small mistake, but in 1939 he published a clear argument which indicated that any intrinsic inconsistencies of quantum electrodynamics were many orders of magnitude away from the experimentally observable domain.

Altogether, in the 1930's the stage was set for the further development of the theory, and a concerted effort of physicists and mathematicians was not completely out of the question. There was no single overwhelming obstacle. Still, the effort would have demanded an unlikely amount of patience and persistence against many stumbling blocks. Quantum electrodynamics has the typical difficulties of a gauge theory, and mathematics was not quite ready to provide elegant tools for their resolution. For the physicists, experiments had not yet provided a compelling reason to put much effort in the study of the small quantum electrodynamical effects of higher order, and to some extent the bewildering features of the other interactions undermined the faith in quantum field theory as a whole. The mathematicians had no reason to invest much work in something which might not last. They still were busy to consolidate the advances of quantum mechanics and gravity theory. Above all, they saw no compelling internal mathematical reason to develop quantum field theory. In hindsight, v. Neumann's operator algebras came close, but they hardly became part of the mathematical mainstream, and v. Neumann soon had more important things to do.

Since the time for quantum electrodynamics was not yet ripe, joint mathematical and physical progress would have needed another stroke of genius. One possibility would have been the creation of a rigorously solvable but non-free toy model, and from today's point of view conformally invariant theories in two dimensions were by far the best thing to be tried. Indeed, there was a little chance. Einstein had thought about conformal invariance, and Dirac got interested in 1936. At that time, Heisenberg just had started to work on quantum field theories with four-fermion interactions. These can be made conformally invariant in two dimensions, such that a joint effort might have led directly to the Thirring model and perhaps to its solution. After the war, Gürsey searched for a way to make Heisenberg's four-fermion theory conformally invariant, in the line of thought which went back to Cunningham and Bateman. He didn't think about two dimensions, however, and wrote down a four-dimensional version with a cube root, which is impossible to quantize [Gürsey 1956]. For Dirac and Heisenberg, it is unlikely, too, that they considered playing around in two dimensions. Moreover, Dirac became increasingly discontent with quantum field theory as a whole.

Many of Heisenberg's efforts were still creative and successful, but his flirt with mathematics was over. Still, his 1932 concept of a nucleon with two states, put four years later in the language of  $SU(2)$  invariance by Cassen and Condon, had initiated the group theoretic studies which from the 50's onward became one of the major themes of particle physics. Here was perhaps a better chance for joint work with mathematicians, for which such considerations soon became very natural. Some physicists also looked at related structures, even before the experimentalists found convincing reasons to study internal symmetry groups or gauge symmetries.

Einstein had long decided to concentrate on classical gravity and electromagnetism and kept away from quantum theory and the nuclear interactions. He continued to work in the context of classical differential geometry and played around with five dimensions and connections with torsion.

These efforts are not highly regarded nowadays, but they familiarized the physics community with the work of E. Cartan and gave much support to the Kaluza-Klein ideas of five-dimensional spacetime. Yang argued that Einstein somehow was looking for the gauge theory found in 1954 by him and Mills [Yang 1982]. Indeed, Einstein repeatedly contemplated parallel transport without the metric constraint of the Levi-Civita connection.

In the five-dimensional line of research, O. Klein himself performed an amazing miracle by writing down the Lagrangian of  $SU(2)$  gauge theory during a 1938 conference in Warsaw. He only recognized the  $U(1)$  part of the symmetry and saw no clear physical applications, since charged vector mesons had not been found yet [Klein 1939, p. 93].

In 1953, Pauli rediscovered the same  $SU(2)$  gauge theory in a conceptually clearer way, when he pushed the Kaluza-Klein ideas one dimension higher and compactified two dimensions on  $S^2$ , such that the  $SO(3)$  symmetry became manifest. Pauli liked the result and described it in a letter to Pais. He did not publish, however, because he saw no mechanism to give mass to the gauge bosons. Together with Heisenberg he started to work on a fermionic Lagrangian with a four-fermion interaction, but he quickly saw that it made not much sense. Cut down from four to two dimensions it would have been transformed from a wrong unified theory to a fascinating mathematical toy. Altogether, Pauli and Weyl were probably the only ones of the pioneering giants who were both close to the mainstream and imaginative enough to push quantum field theory by inventing, e.g., a mathematically nice conformal field theory. In more fortunate times, Zürich might have witnessed such a step ahead, but it is hard to play in the shadow of war and persecution.

With roots in a dominating wave of mood, the Nazi aversion against Jewish mathematics and physics had pervaded the German universities and the Göttingen environment was destroyed. The focus of research shifted from Europe to the USA.

### 3 Progress in the face of mathematics

After the war, the seminal event in the further development of quantum field theory was the Shelter Island conference. The decisive input came from the experimentalists, who made good use of the technology created in the years before. Their results implied that quantum electrodynamics had to be taken very seriously. Mathematicians were absent. Apparently, it had occurred to nobody that they might be of help.

It seems that progress needed a new generation: Very attentive to the experiments, much less to new mathematical structures, conservative in its attachment to the old principles found in the golden age, careful and innovative in calculations. In Princeton, the lonely walks of Einstein and Gödel were parts of a different world, faint reverberations of revolutions in a distant past. The young mathematicians had happy times. They developed fibre bundles, connections, characteristic classes, the deformation theory of complex structures and many other nice things. They laid the groundwork for modern physics, and couldn't care less.

In 1947, work on interacting quantum fields started in earnest. Bethe explained the Lamb shift, and soon after Schwinger calculated the anomalous magnetic moment of the electron. The calculations still were done with the theoretical tools of the prewar period. They started from the quantum theory of free fields and introduced perturbations according to the standard rules of quantum mechanics. Since quantum mechanical perturbation theory makes no use of Lorentz invariance, this procedure compounded the intrinsic difficulties. Soon after, Schwinger and Feynman developed relativistically invariant formalisms, and comparison of quantum electrodynamics with the experiments became very successful. Stueckelberg and Tomonaga had done earlier work in this direction, unfortunately with less impact.

These calculational procedures were correct, but were derived from wrong standard assumptions by dubious mathematical methods. Soon, the standard assumptions were proven to be wrong by

a small group of mathematical physicists, whose work was based on an uncontested set of axioms. This caused some uneasiness among the calculators, and kept the mathematical community at a distance. To explain what happened, we have to consider some details of the Heisenberg and Pauli paper of 1929. They were the first to derive the equal time commutators of free fields.

With respect to differentiation by space and time coordinates, quantum fields can satisfy differential equations. In the simplest case, the latter are linear, as for Maxwell's equation. Such fields are called free, since a linear combination of two solutions describes two waves which pass through each other without mutual influence. In the language of today, one starts with the space of classical solutions of the linear differential equation. On this space one needs a symplectic structure, given by a Poisson bracket, or equivalently a Heisenberg Lie algebra. In most cases one has an invariance with respect to a time translation group, the generator of which is the energy. Polarization with respect to the sign of the energy yields the appropriate Hilbert space representation of the Heisenberg Lie algebra (Fock space).

Apparently, in the 20's the symplectic structure of the space of classical solutions had not yet been grasped in depth. Thus the historical procedure was slightly more complicated and involved a special and somewhat formal choice of the classical observables. For free fields, it yielded canonical commutation relations in perfect analogy with Heisenberg's commutation relations

$$[x_i, p_j] = i\delta_{ij} ,$$

and the vanishing equal time commutators  $[x_i, x_j]$  and  $[p_i, p_j]$ . Analogously, for a free real scalar field  $\phi$  satisfying the Laplace differential equation, the equal time commutators  $[\phi(x), \phi(y)]$  and  $[\partial_t \phi(x), \partial_t \phi(y)]$  both are zero. The remaining equal time commutator takes the form

$$[\phi(x), \partial_t \phi(y)] = i\delta(x - y) .$$

Note that distribution theory was not yet developed. This caused no physical problems at all, but meant that the use of Dirac's  $\delta$  had no firm mathematical base yet. One may wonder, if this encouraged the physics community to ignore mathematical niceties. Of course, the mathematical justification was provided in the late 40's by L. Schwartz, in a splendid case of interaction between the two communities. When this was done, one had a good description for free quantum fields as distributions over three-dimensional space. Once a fixed field  $\phi$  is paired with a test function  $f$  (physicists write  $\int \phi(x) f(x) d^3x$ ), the result is an element of the Heisenberg Lie algebra and acts on the Hilbert space of the system. For real  $f$  the operator is hermitean and describes an observable, as usual in quantum mechanics.

The analogy between particles and free fields given by passing from the Kronecker function  $\delta_{ij}$  to Dirac's  $\delta(x - y)$  was compelling, but proved to be very misleading. Heisenberg's commutation relations for  $x_i, p_i$  remain valid when interactions are present. In contrast, Haag showed that an interacting field theory cannot have canonical commutation relations. Indeed, interacting fields cannot even be understood as distributions over three-dimensional space at fixed time. Time averaging is necessary, too [Haag 1955].

In a special relativistic context, this might not have come as a big surprise. There even was a paper by Bohr and Rosenfeld which argued that a careful analysis of measurements implies a spacetime average [1933]. The arguments were clear enough and the paper was never forgotten, but its somewhat obscure style missed its mark on most of the new generation.

Instead, adherence to the canonical commutation relations for quantum fields remained pervasive in the physics literature till recent times, inspite of the fact that everyone knew it was wrong. Most probably, it was much more this attitude than the difficulties of renormalization which made it impossible for mathematicians to digest the intricate and important structures of quantum field theory.

Historians will have to weigh this issue when the dust has settled. Despite of what has been said, they hardly can find a better starting point than the following classic quotation: "In the thirties, under the demoralizing influence of quantum-theoretic perturbation theory, the mathematics required of a theoretical physicist was reduced to a rudimentary knowledge of the Latin and Greek alphabets." (Jost) [Streater and Wightman 1964, p. 31].

The insistence of the physics community on using a wrong basis for successful calculations would be easy to understand, if no alternative formalism had been available. Due to Schwinger and Dyson, this was not the case. Dyson had read much mathematics and brought clarity of thinking to the muddled field. By 1949, Schwinger and Dyson had started to analyse quantum fields in terms of the  $n$ -point functions (or rather distributions)  $T\langle\phi(x_1, t_1) \dots \phi(x_n, t_n)\rangle$ . Here for an operator  $A$  the real or complex number  $\langle A \rangle$  is its expectation value in the vacuum state of the Hilbert space, and the analogous notation applies to distributions. The time ordering imposes the condition  $t_1 > \dots > t_n$  on the support of the test functions. Moreover, in 1951 Schwinger published his action principle, which describes how an  $n$ -point function varies when one changes the parameters of the interaction.

Thus most of the theoretical tools were ready. On reading the tributes to Schwinger published after his death [Ng 1996], it seems that some obstacles to progress were personal. Schwinger had been a prodigy and the centre of attention. Apparently, he didn't mind that his calculations remained almost incomprehensible. All that changed after 1948. In Schwinger's own words: "Like the silicon chip of more recent years, the Feynman diagram was bringing computation to the masses" [Schwinger 1983, p. 343]. Dyson had a particularly clear understanding of the issues: "The advantages of the Feynman theory are simplicity and ease of application, while those of Tomonaga-Schwinger are generality and theoretical completeness" [Dyson 1949, p. 486]. Schwinger forbade his students to mention Feynman or Dyson, or to use Feynman graphs. From a European perspective it seems that Einstein and Weyl would have had more reasons for grudges against Hilbert and Schrödinger, but one has to respect a difference of culture.

In 1953, the Wightman axioms [Streater, Wightman 1964] were presented in lectures at Princeton. They were something of a mixed blessing. On one hand, they allowed clear proofs of structural statements, in particular of Haag's insight that the canonical commutation relations are wrong for interacting theories [Haag 1955]. On the other hand, the axioms sacrificed the connection to the concrete quantum field theories which were under development.

One technical detail needs comment. The Wightman axioms concern  $n$ -point distributions  $\langle\phi(x_1, t_1) \dots \phi(x_n, t_n)\rangle$ , but without time ordering. This seems mathematically convenient, for example when one wants to take Fourier transforms. Nevertheless, for contact with the experiments, the time ordering is natural. This became particularly clear with the LSZ formalism of Lehmann, Symanzik and Zimmermann, which provided a direct calculation of the results of scattering experiments in terms of the time ordered distributions. Different time orderings correspond to different experiments.

The three authors were members of Heisenberg's group, which attracted most of the young people who wanted to work on elementary particles in postwar Germany. Unfortunately, Heisenberg was hardly interested in mathematics and too occupied by his world formula to have much regard for the LSZ achievements. When Lehmann returned from the States, Heisenberg greeted him: "Na, Herr Lehmann, wie geht's der Mathematik?" (how is mathematics?), an episode which Lehmann never forgot. So much for the superiority of European culture.

As an aside, any Third World country which wants to strengthen her scientific basis would be well advised to do a few case studies. The decline of physics in Germany is particularly interesting. One cannot put all of the blame on fascism, since mathematics did not suffer the same fate after the war, largely due to the achievements of Hirzebruch.

The  $n$ -point distributions made mathematical sense, but were difficult to deal with. The next big advance was the introduction of the euclidean formalism, as discussed in [Osterwalder 1973]. Early on, Dyson had recognized that some calculations become much easier when one performs

an analytic continuation to imaginary values of time (Wick rotation). The gestation of the idea took most of the 1950's, with contributions from Wick, Nakano and, in condensed matter physics, Matsubara [1955]. It first appears in complete form in papers of Schwinger.

In his 1993 lecture in Nottingham [Ng 1996], Schwinger states that it could have been published any time after 1951, but in fact "The Euclidean Structure of Relativistic Field Theory" appeared in 1958. Schwinger made an analytic continuation of the time-ordered  $n$ -point distributions to purely imaginary values of time. As Wightman had seen already, the analytic continuation allows to consider the distributions as boundary values of ordinary analytic functions. Thus Schwinger's idea allows to describe physics by functions of some  $D$ -dimensional euclidean space instead of distributions with testfunctions over  $D$ -dimensional spacetime. By that time, mathematical physicists had mastered the difficulties of distribution theory, such that the due expression of relief was rather muted. Often, the euclidean  $n$ -point functions are regarded as distributions, too, but the present article will not follow this habit.

As usual nowadays, Schwinger's euclidean  $n$ -point functions will just be written in the form  $\langle \phi(x_1) \dots \phi(x_n) \rangle$ , where the  $x_i$  now denote points in  $D$ -dimensional euclidean space. These functions are real analytic and defined everywhere except on the partial diagonals  $x_i = x_j$ . Since there is no causal structure in euclidean space, the necessity of time ordering disappears. Accordingly, the functions are symmetric under permutation of the  $x_i$ . If one considers several fields  $\phi_1, \phi_2, \dots$ , one has instead

$$\langle A_1 \phi_i(x_i) \phi_{i+1}(x_{i+1}) A_2 \rangle = \langle A_1 \phi_{i+1}(x_{i+1}) \phi_i(x_i) A_2 \rangle ,$$

where the  $A_k$  stand for products of fields at points different from  $x_i, x_{i+1}$ . In spacetime, all possible time orderings can be reached by analytic continuations starting from the same euclidean  $n$ -point function, a fact called crossing symmetry.

Since the choice of quantum field theories is quite limited, their  $n$ -point functions should be special functions with very interesting properties. Not much is known about them, however. For free theories, they vanish unless  $n$  is even, in which case they reduce to sums of products over 2-point functions. The latter are variants of Bessel functions. For conformal field theories, one obtains functions of hypergeometric type. In some other cases in two dimensions, at least the 2-point functions are under good numerical control, but little is known about their analytic properties. It is quite possible that some examples will yield functions of Painlevé type. Unfortunately, interest in special functions was at a low ebb in the past century, but this certainly will change again.

Most quantum field theories have free parameters. The latter take values in some differentiable manifold which is called moduli space. Accordingly, the  $n$ -point functions can be differentiated with respect to these parameters. Let  $\partial_\lambda$  be a tangent vector in moduli space. According to Schwinger's action principle, each tangent vector corresponds to some field  $t(x)$ , such that formally

$$\partial_\lambda \langle \phi(x_1) \dots \phi(x_n) \rangle = \int \langle t(x) \phi(x_1) \dots \phi(x_n) \rangle d^D x .$$

The expression is formal, since the integral diverges when  $x$  approaches one of the  $x_i$  and needs to be regularized.

In general, there is no easy way to normalize the field  $\phi$ . Of course, the canonical commutation relations would have provided a natural normalization, but they are wrong. When one changes the normalization by some factor  $f(\lambda)$ , the derivative of the  $n$ -point function changes by a term proportional to  $n \langle \phi(x_1) \dots \phi(x_n) \rangle$ . If the divergence of Schwinger's integral is of exactly this type, the freedom of normalization can be used to cancel it. This is the renormalization procedure, which will be discussed in more generality below.

In principle, vector fields can be integrated, such that Schwinger's action principle should allow to recover the moduli space from any of its regular points by higher order derivatives and the summation of the Taylor expansion. In many practical cases, however, the only explicitly known points of the moduli space lie at the boundary, where the space is no longer regular. As a



consequence, the Taylor expansion is only asymptotic. This problem can be avoided for conformal field theories, but it will be mentioned again in the context of string theory.

Many moduli spaces do not have a natural metric, such that the integration of a vector field has to follow an arbitrary smooth curve. Equivalently, one can choose local coordinates, also known as renormalization scheme. Indeed, without a metric on moduli space, the perturbing field  $t(x)$  does not have a natural normalization. Typically it lives in some finite dimensional vector bundle over moduli space which includes mass perturbations and coupling constant perturbations. When one takes higher order derivatives of the  $n$ -point functions, all of these parameters have to be considered together, which requires mass and coupling constant renormalizations of  $t(x)$ . The finite ambiguities of the latter are fixed by the choice of a renormalization scheme. Changing them leads to a different curve for the integration.

If one wants, one can include the constant field 1 in the vector bundle, but since one wants  $\langle 1 \rangle = 1$  it is usually more convenient to require  $\langle t(x) \rangle = 0$ . This is called the renormalization of the vacuum energy density.

In the 50's, renormalization was well understood on a computational level, but before Wilson's work in the late 60's the concepts were not particularly clear. Nevertheless, the time was ripe for the first quantum field theory which was not free and made complete mathematical sense.

## 4 Thirring model: Conformally Invariant QFT Is Born

In 1958, W. Thirring published a paper with the title 'A Soluble Relativistic Field Theory' (in Mathematical Reviews, it was described by Raychaudhury, Calcutta). The paper kept the promise of its title. Let me quote a few sentences: 'In spite of the great efforts of many people the mathematical structure of relativistic quantum fields is still in the dark. ... In order to study those (features) we propose in the present paper a model of a relativistic field theory... Since the reduction of the number of fields does not simplify the problem sufficiently ... one has to take recourse to a reduction of the dimensionality of the problem... Thus the simplest nontrivial case seems to be a one-dimensional Fermi-field with an interaction  $\lambda \bar{\psi} \psi \bar{\psi} \psi$ . Although the problem is of considerable complexity it turns out to be soluble. ... (The model) shows explicitly what a relativistic theory can look like. Furthermore it can serve as a testing ground for field theorists.'

All of this is true. Perhaps the most remarkable part is the courage to do something simple in two dimensions. Here Thirring was inspired by the investigation of many-body systems in terms of the Bethe ansatz. In two dimensions, one can get interactions by collisions only, without fields. This knowledge led to the correct conjecture that the model would be solvable. Thirring also made some entirely correct remarks about Heisenberg's unified four-fermion interaction theory in four-dimensional spacetime, which may have contributed to some tension between Munich and Vienna. Indeed, despite of the fact that part of Thirring's work had been done at MIT and at the IAS, Princeton, one almost gets the impression that the creation of the model was a provincial non-event. The leading soluble model of the time was due to Lee (1954) and not relativistic. Thirring's remark about the Lee model in his 1958 paper is not particularly deferential, but in his textbook with Henley [1962] he gives it two chapters, whereas his own model does not even seem to be hinted at. Schweber's 1964 textbook doesn't cite it either.

Nevertheless, some of Schwinger's former students had paid attention, and Johnson from MIT devoted a paper to the model [1961]. I quote from the introduction: "Thirring has proposed a two dimensional ... model which is of some interest because its exact solubility enables one to study some of the general conjectures which have been proposed in regard to the behaviour of local relativistic fields. In spite of the model, no general solutions have been proposed which are free from possible criticism because of the rather formal manner in which they have been obtained." In the conclusion, Johnson states: "We have shown how it is possible to solve the two dimensional



model of Thirring by making use of the existence of the two vector density conservation laws. ... It was shown how it is possible to define the products of the singular operators  $\psi(x)$ , in order to determine other covariant operators but that these singular field products do not satisfy the equal time commutation relations with the field operators  $\psi(x)$ , that one would obtain by means of the canonical commutation relations ...". Again, all of this is correct. Still, some mathematical problems were left, but they were settled in the subsequent years.

Let us describe the model in more detail. It is obtained by perturbing the theory of a massless complex fermion in two dimensions. In the euclidean formulation, the Dirac equation reduces to the Cauchy-Riemann equation and its complex conjugate. Real and imaginary parts of the fermion yield two holomorphic field  $\psi_i(z)$  and two anti-holomorphic fields  $\bar{\psi}_i(\bar{z})$ ,  $i = 1, 2$ . At this point in moduli space, the two conserved vector densities mentioned by Johnson are  $j(z) = \psi_1(z)\psi_2(z)$  and  $\bar{j}(\bar{z}) = \bar{\psi}_1(\bar{z})\bar{\psi}_2(\bar{z})$ . The conservation equations are the Cauchy-Riemann equation for  $j$  and its conjugate for  $\bar{j}$ . The 2-point functions have the form

$$\langle j(z_1)j(z_2) \rangle = (z_1 - z_2)^{-2}$$

and analogously for  $\langle \bar{j}\bar{j} \rangle$ , whereas  $\langle j\bar{j} \rangle = 0$ .

In terms of Schwinger's action principle, the perturbation corresponds to the field  $t = j\bar{j}$ . It turns out that the  $n$ -point functions of  $j$  and  $\bar{j}$  are unaffected by the perturbation. In particular, the two fields and their product  $t$  have a natural continuation over Thirring's moduli space and need no renormalization. Moreover, the conservation equations do not change, which accounts for the solvability of the model.

The special properties of  $j, \bar{j}$  arise because they are currents, i.e. quantum analogues of the conserved densities which arise by Noether's theorem from continuous symmetries. Because of their close relation to observable quantities, they behave similarly to free fields. This led to the concept of current algebra. In two dimensional theories, the currents of simple Lie groups generate the corresponding affine Kac-Moody algebra, at least when space is compactified to a circle. Unfortunately, the mathematical potential of current algebras was not realized for many years. The work of Kac and Moody in 1967 was independent of physics. In the context of string theory, it was introduced in the physics literature by the mathematicians Lepowsky and Wilson [1978] and again by G. Segal [1981], and became a rare success story of physics and mathematics in cooperation.

The Thirring model fields  $\psi_i$  do not remain holomorphic under the perturbation by  $j\bar{j}$ . Instead, one obtains

$$\langle \psi_i(z_1)\psi_j(z_2) \rangle = (z_1 - z_2)^{-1}|z_1 - z_2|^{-s}\delta_{ij} ,$$

where the real number  $s$  changes under perturbation. Under the conformal transformation  $z \mapsto z' = (az + b)/(cz + d)$  with  $ad - bc = 1$ ,  $\psi(z) \mapsto (cz + d)^{-1}|cz + d|^{-s}\psi(z')$ , the two-point functions remain invariant. This remains true for all the  $n$ -point functions, such that the Thirring model is a conformally invariant theory. Initially, this seems to have been overlooked, and only the special case of invariance under scale transformations  $z \mapsto \lambda z$  was commented upon. This is a bit surprising, since in these years Thirring was very much concerned with conformal invariance. In the important 1962 paper where Gell-Mann introduced current algebra to the theory of the strong interactions, he acknowledges that Thirring introduced him to the conformal group. Moreover, conformal invariance had become an issue between Munich and Vienna. There was little internal logic in this local turbulence, but it turned out to be important and may be of interest to historically inclined people.

Heisenberg had developed an interacting spinor theory for all of particle physics and pushed it for many years, though it made no sense. At the time, the new quantum number of strangeness demanded an explanation. Due to Noether, an invariance of the theory had to be found. Heisenberg tried scale invariance, though the theory has a length scale and a non-compact group has a hard time to yield discrete quantum numbers. The contemporary fashion for negative norm states, also present in the Lee model, gave some hope for a cure [Dürr 1959].

In Vienna, Cunningham and Bateman were remembered and Wess used Heisenberg's attempts as justification for the resurrection of the conformal group. In a brief remark, he hinted at a possible use of the conformal group at high energies. Otherwise, he showed in a few pages that Heisenberg had missed the mark [Wess 1960]. Since several of the few good young German theoreticians had flocked around Heisenberg, the paper triggered new interest in the conformal group, and Kastrup started to work on it, though Heisenberg did not pay much attention. Kastrup published papers on the possible importance of conformal invariance at high energies. During a visit to Russia, he explained it to Polyakov, as acknowledged in the first paper of the latter on conformal symmetry [1970]. This paper showed that scale invariance implies full conformal invariance. On the other side of the Atlantic, in his historic paper on the short distance expansion, Wilson ascribes the idea of scale invariance at short distance to Kastrup and his student Mack [Wilson 1969]. The fact that scale invariance implies full conformal invariance was recognized by Callan, Coleman and Jackiw, slightly before Polyakov's work and in a different context [1969]. On the physical relevance of scale and conformal invariance, they cite 1969 papers by Mack and Salam and by Gross and Wess.

Wilson's short distance expansion was the main concept which still was lacking for a rigorous and computationally efficient description of quantum field theory. It concerns the behaviour of the  $n$ -point functions along their singularities. Wilson considered them in Minkowskian spacetime, but the euclidean case is much easier.

It has been mentioned that the euclidean  $n$ -point functions  $\langle \phi(x_1) \dots \phi(x_n) \rangle$  are not well defined on the partial diagonal  $x_i = x_j$ . In general, the functions diverge on these diagonals. For a free field  $\phi$  of dimension  $h$ , the leading term at  $x_1 = x_2$  is proportional to  $|x_1 - x_2|^{-2h} \langle \phi(x_3) \dots \phi(x_n) \rangle$ . The case of several different fields needs a bit more discussion, but is not complicated either.

There had been some speculation on the corresponding behaviour for interacting fields. One idea was that the singularity might be the same as for free fields. In 1964 Wilson conjectured that perturbations just introduce some logarithmic corrections. This was wrong, but one of Wilson's talents was to talk to the right people for correcting mistakes. In particular, he had crucial discussions with Johnson, who familiarized him with the Thirring model. Wilson learned that the latter indeed is scale invariant, but that the dimension  $h$  changes with the strength of the interaction. Independently, the same modification to Wilson's original ideas was made by Lowenstein.

Wilson was a mainstream theorist on the way to a Nobel prize, but he did not fear to go against the tide: "The assumption that integrating an operator over space only gives an observable is a basic tenet of canonical field theory... The assumption has been rejected by axiomatic field theory from the beginning" [Wilson 1970, p. 1484]. In the same paper, he discusses a related issue and concludes: "The axiomatic view must in the end replace the popular view" [p. 1483]. It seems that the time was ripe to discuss all of quantum field theory in terms of statements which are at least potentially true.

Before we discuss other aspects of Wilson's work, let us continue the history of the Thirring model. At the end of the 60's, string theory was invented and soon it was recognized that conformal field theory is an essential ingredient [Galli 1970]. Halpern recognized the importance of the Thirring model in this context and informed Virasoro, who gave it publicity [1971]. A comparative investigation of the Thirring model and string physics in the context of conformal field theory was made by Ferrara, Grillo and Gatto [1972].

By 1974, it had become popular to elucidate the properties of quantum field theory by a study of two-dimensional examples. A particularly interesting one was the sine-Gordon model, which describes a bosonic scalar field with trigonometric interaction term. Coleman wrote an elegant and deep paper where he showed that the perturbation by a fermion mass term makes the Thirring model isomorphic to the sine-Gordon model [1975]. This took everyone by surprise, since superficially the two models look entirely different and equally impenetrable in a strict mathematical sense.

On hindsight, people remembered that the equivalence between fermions and bosons in two dimensions had been prefigured by Skyrme [1958, 1961], but Skyrme had been too far ahead to

have an immediate impact.

After Coleman's paper, at last, one leading mathematician was shocked enough to take things seriously. G. Segal regarded the mass term as an unessential complication and concentrated on the boson-fermion equivalence. This was Coleman's starting point and concerns an isomorphism between two conformally invariant theories. Initially, Segal felt quite sure that boson-fermion equivalence made no sense. When it turned out in the late 70's that the equivalence leads to a combinatorial identity known already to Euler, a dam had been broken. Segal developed a beautiful system of axioms for conformally invariant quantum field theories in two dimensions and transformed the latter into a legitimate field of study for mathematicians [Segal 1988]. But even in their book on loop groups [Pressley, Segal 1986, p. 215] the authors state that a mathematically clear formulation of the isomorphism between the massive Thirring model and the sine-Gordon model still seems not to have been found.

## 5 Nature's helping hand

The long delay in the gestation of a correct theory of quantum fields would have been even longer without some direct help from nature. One reason is that the investigation of two-dimensional toy models was not taken very seriously by the particle physicists. Here is a quotation from a paper which reports the discovery of a fundamental property of the Thirring model: "The results are of interest ... because they allow one to see very readily (a) why the Thirring model is solvable and (b) why it has trivial physical consequences. As will be clear from the following, the solvability of this model depends critically on the fact that it is a 2-dimensional model. It is not likely that any of the specific features of this model can be generalized to more realistic cases, or that they will provide a useful guide to the state of affairs in the real world" [Callan, Dashen and Sharp 1968, p. 1883].

Indeed, the highly non-trivial physical consequences of such conformal field theories in the context of string theory could not have been guessed in 1967. No wonder that the authors permitted themselves some sloppiness in the analysis: "At this point, one could introduce the Fock representation for the scalar field, annihilation and creation operators, etc., and verify in detail that the energy and momentum operators have the expected properties, but there is little to be gained by going over these well-known details" [p. 1885]. This was a missed opportunity. For example, the commutation relations for the energy-momentum tensor given in the paper miss the central extension of what is now called the Virasoro algebra. What would have happened, if some interested mathematics student had tried to digest the paper?

Since it seems that no mathematicians were interested, it was very kind of nature to provide her own motivation for the study of such models. In the 50's, physicists were confronted unexpectedly with a rich class of quantum field theory in condensed matter laboratories, which turned out to be conformal field theories in the real world of two-dimensional surface coatings or three dimensional liquids.

After Feynman's breakthrough in 1948, his graph methods soon were transferred to other fields of physics. Their application in condensed matter physics was pioneered by Salam [1953] and Matsubara [1955]. In particular, Matsubara recognized the perfect analogy of imaginary time and temperature, due to the relation between the time translation  $\exp(iHt)$  in quantum mechanics and the Boltzmann factor  $\exp(-H/T)$  in statistical mechanics.

When continuous phase transitions were studied, it turned out that the analogies went much deeper. At the critical temperatures, the behaviour of the materials is dominated by long range fluctuations of arbitrary scales, and the details of the molecular structure become unimportant. The theory approaches a continuum limit. The correlation functions of the limiting theory behave exactly like the euclidean  $n$ -point functions of quantum field theory. In this way, many statistical

systems at continuous phase transitions are related to quantum field theories in spacetime by analytic continuation.

Thus nature herself had declared that the Wick rotation introduced by Schwinger makes good sense. Of course, the dimensions of the observed examples are different, since the phase transitions happen in two or three dimensional systems, whereas spacetime has four dimensions. Moreover, the natural constraints on the field theories are not the same. Quantum field theories need a probability interpretation, which is realized by positive scalar products. Under Wick rotation, this becomes Osterwalder-Schrader positivity, which is not a necessary property of phase transitions. On the other hand, statistical observables are given by real numbers. This real structure yields a time reversal invariance of the corresponding quantum field theory, a property not shared by all examples and only approximately true in nature. On a purely mathematical level, these difficulties are not particularly serious, however.

Lab experiments on phase transitions were much cheaper than particle physics with high energy accelerators. Moreover, there were no worries that a breakthrough in the domain of the fundamental laws was necessary. Thus progress was rather steady, both on the experimental and the theoretical side. Soon it became clear that the physics at the critical phase transition point is scale invariant [Kadanoff 1966]. Much of the relevant work on these euclidean quantum field theories was done in the Soviet Union, and Polyakov was one of the most important contributors. He found convincing arguments that scale invariance implies full conformal invariance at the critical point and recognized that this invariance allowed a calculation of the 3-point functions up to a constant factor [Polyakov 1970].

Further developments depended on the analysis of a soluble example in the context of statistical mechanics. This might have been provided by the Thirring model, which had occurred in its bosonic description, and was called the gaussian model. Because of the somewhat misleading simplicity of the bosonic formulation, the subtle features of its fermionic fields were not recognized in this context. Instead, the Ising model played a rôle for the study of continuous phase transitions which was parallel to the one of the Thirring model for particle physicists.

The states of the Ising model put a number 1 or -1 to each site of a rectangular lattice. The latter are called values of the Ising spin. Pairs of nearest neighbours have an interaction energy which depends on the product of their Ising spins. The total energy  $E$  is given by a sum over the interaction energies of such pairs. The thermodynamic partition functions at temperature  $T$  is given by the average of  $\exp(-E/T)$  over all states.

Rectangular lattices can be considered in various dimensions. The thermodynamic functions for the linear or one dimensional model are very easy to calculate. The problem was given by Lenz as part of a PhD thesis to a rather weak student, who did not do any later scientific work. One hardly can imagine an easier way to lasting fame. The two-dimensional model, where the Ising spins sit on a square lattice, was solvable but very hard. The breakthrough calculation was due to Onsager [1944]. There is a critical temperature where the model turns into a rather simple euclidean quantum field theory. In particular, at this point the spin waves of the model satisfy the two-dimensional Dirac equation for free massless fermions, as first noted by Kadanoff [1969]. This equation is conformally invariant, as in the more complicated four-dimensional situation. In contrast to the complex fermion of the Thirring model, the fermion field of the Ising model is real. In this sense, the Ising model at its critical temperature has half as many degrees of freedom as the Thirring model.

The two-dimensional Ising model is not just a theory of free fermions, however. The average values of the Ising spins turn into a field with scaling dimension  $1/8$ . This result proved to be a highly non-trivial check which uncovered the failures of many calculational methods.

Now two different conformally invariant quantum field theories were available, the Ising model in statistical mechanics and the Thirring model in conventional relativistic quantum field theory. They were used for very much the same theoretical tools, in particular the short distance expansion. Wilson discovered it in 1964 in the Minkowskian context, Polyakov and Kadanoff in 1969 in

the euclidean. Polyakov called it correlation coalescence, Kadanoff reduction hypothesis. Wilson called it operator product expansion, and this terminology has survived, because it clearly has the priority. In the context of statistical mechanics it is not appropriate, however, since there are no operators around. Since it has advantages to have a unique name in both contexts, we use the common synonym short distance expansion.

A scale invariant  $n$ -point function of type  $\langle \phi(x)\phi(y)A \rangle$  has a leading singularity at  $x = y$  proportional to  $|x - y|^{-2h}$ , where  $h$  is the scaling dimension of  $\phi$ . When this leading singularity is subtracted, the next term behaves like  $|x - y|^{-2h+h_1} \langle \chi_1(y)A \rangle$ . Here  $\chi_1$  is some other field of scaling dimension  $h_1 > 0$ , which can be measured in the way just described. Subtracting this subleading term one finds  $|x - y|^{-2h+h_2} \langle \chi_2(y)A \rangle$ , where  $\chi_2$  now has a larger scaling dimension  $h_2 > h_1$ . The procedure can be repeated as far as one wants to go. One will find an infinity of fields of ever higher scaling dimension. Note that the  $\chi_i$  are independent of the fields included in  $A$ .

One can apply the same procedure to other  $n$ -point functions like  $\langle \phi(x_1)\chi_1(x_2)\dots \rangle$  and so on and produce as many new fields as possible. The short distance expansion now states that for any real number  $h_0$  there is only a finite number of linearly independent fields of scaling dimensions  $\leq h_0$ . This property can be verified in many concrete examples and may very well be taken as part of the mathematical definition of a quantum field theory.

Lattice systems are scale invariant at the exact temperature of a continuous phase transition. When the temperature is changed a bit, the correlations will show an exponential decay at large distances. When one is sufficiently close to the critical temperature, the corresponding correlation length is still very large compared to the distance between neighbours. With a suitable limiting procedure, one obtains the  $n$ -point functions of a quantum field theory which is no longer conformally invariant. In this case, more complicated expressions than  $|x - y|^{-2h}$  will occur in the  $n$ -point functions. At the very least, one expects logarithmic correction factors. Nevertheless, the basic idea of the short distance expansion applies as before.

Let us consider a euclidean  $n$ -point function  $\langle \phi(x)\chi(y)A \rangle$ , where  $A$  is a product of local fields at positions different from  $x, y$ . An experimentalist may study the behaviour of this function when  $x$  approaches  $y$ . Each such measurement can be interpreted as the measurement of some field at  $y$ . This is the physical content of the short distance expansion. We can axiomatize it in the following way. Let  $\Gamma(y)$  be the vector space of germs of functions which are defined near  $y$ , but not at the point  $y$  itself. We give a topology to this space by using  $o(|x - y|^s)$ ,  $s \in \mathbf{R}$  as a basis of neighborhoods of 0 in  $\Gamma(y)$ . Let  $\gamma$  be an element of the dual of  $\Gamma(y)$ . Then for each pair of fields  $\phi, \chi$  and each  $h$  there must be a field  $\psi$  such that  $\gamma(\phi(x)\chi(y)A) = \langle \psi(y)A \rangle$  for arbitrary  $A$ . One just can write  $\gamma(\phi(x), \chi(y)) = \psi(y)$ .

Consider the vector space  $F$  of all fields of a quantum field theory. This vector space is filtered by the scaling dimension. Let  $F(h)$  be the subspace of all fields of scaling dimension less or equal to  $h$ . We assume that these subspaces are finite dimensional. We also assume that the theory has some degree of asymptotic scale invariance. More precisely,  $\psi \in F(h_1 + h_2 + h_0)$  when  $\phi \in F(h_1)$ ,  $\chi \in F(h_2)$  and  $\gamma$  vanishes on  $o(|x - y|^h)$  for  $h > h_0$ . This condition will be important for renormalizability. Finally,  $\dim F(h)$  should not increase faster than for free theories. In two dimensions, this yields  $\log(\dim F(h)) = O(\sqrt{h})$ .

In this way one obtains a nice algebraic structure which is well adapted to calculational purposes. It does not contradict the Wightman axioms, but emphasizes quite different aspects. Whereas those axioms concentrate on one field, or maybe a few, the short distance expansion considers all possible fields at once. For mathematicians, this is certainly the more natural procedure. To some extent, it eliminates the surprise one first feels about the equivalence of the sine-Gordon and the massive Thirring model, since in the latter one immediately has to include its bosonic fields, too.

## 6 Regularization and renormalization

With the help of the short distance expansion, it is rather easy to put renormalization in a standard mathematical frame. First we have to generalize the change of normalization of the fields which we considered above. Instead, we will use all the linear transformations of  $F$  which conserve the subspaces  $F(h)$ . The group of these linear transformations will be called  $L(F)$ .

We want to regard a perturbation of some theory. In accordance with Schwinger's action principle, the deformation is described by a field  $t(x)$ . We shall see that in a spacetime of  $D$  dimensions, the scaling dimension of  $t$  must be  $D$  or less.

The corresponding derivative of an  $n$ -point function  $\langle \phi_1(x_1) \dots \phi_n(x_n) \rangle$  is given by  $\int d^D x \langle t(x) \phi_1(x_1) \dots \phi_n(x_n) \rangle$ . The integral behaves well at infinity, but diverges when  $x$  approaches one of the  $x_i$ . Thus we regularize it by excluding a small neighborhood of size  $\epsilon$  around each  $x_i$  from the integration domain. Let us denote the resulting integral by  $\int_\epsilon$ .

The idea of renormalization means that the divergence can be absorbed by a redefinition of the fields. Such a redefinition is given by a linear transformation in  $L(F)$  of the fields which maps every subspace  $F(h)$  into itself. Using Wilson's short distance expansion, one sees easily that there are transformations  $f(\epsilon) \in L(F)$  such that

$$\int_\epsilon d^D x \langle t(x) \phi_1(x_1) \dots \phi_n(x_n) \rangle = \sum_{i=0}^n \langle \phi_1(x_1) \dots (f(\epsilon)\phi_i)(x_i) \dots \rangle$$

has a well defined limit when  $\epsilon$  goes to zero. Indeed, any divergent contribution  $\gamma$  to the integral near  $x_i$  vanishes when the  $n$ -point function behaves as  $o(|x - x_i|^{-D})$ , such that  $\gamma(t(x)\phi) \in F(h_i)$ , when  $h_i$  is the scaling dimension of  $\phi_i$ .

The transformation  $f(\epsilon)$  is only defined up to addition of a finite linear transformation in  $L(F)$ . Any choice defines a connection on the filtered vector bundle  $F$  over the moduli space. Altogether, we now have well defined first derivatives in the moduli space of a quantum field theory. The calculation gets harder when one looks at higher derivatives, since the perturbing field  $t(x)$  will have to be renormalized, too, but this is just a technical difficulty.

As one sees, renormalization is nothing particularly problematic. On the contrary, regularization of divergencies has a long history in mathematics. For example, the Weierstrass product formula for entire function needs the regularization of an infinite product. Let us consider it in more detail. One wants a product formula for an entire holomorphic function  $P(z)$  with zeros exactly at given positions  $z_i$ ,  $i = 1, 2, \dots$ , more precisely a function with  $\sum(z_i)$  as zero divisor. The sequence  $z_i$  must have no accumulation point in the Gauss plane. When the number of zero positions is finite, the product  $\prod(z - z_i)$  will do. The most general function with this divisor is  $\exp(f(z)) \prod(z - z_i)$ , where  $f(z)$  is an arbitrary entire function.

Now let us consider the case of an infinite number of positions. Factoring out a power of  $z$  if necessary, we may assume that none of the  $z_i$  is zero. Let us formulate Weierstrass' solution in terms of the language of quantum field theory. We regularize the problem by restricting the set of zeros to  $z_i$ ,  $i = 1, \dots, N$ . Then we order the  $z_i$  in accordance with their absolute value and renormalize the function  $\prod_{i=1}^N (z - z_i)$  in the form

$$P_N(z) = \exp(f_N(z)) \prod_{i=1}^N (z - z_i),$$

such that the limit  $\lim_{N \rightarrow \infty} P_N$  is finite.

The situation in quantum field theory is quite analogous. The cut-off by  $\epsilon$  is analogous to the cut-off by  $N$ , the achievement of convergence by the renormalization transformation  $f(\epsilon)$  is analogous to the multiplication by  $\exp(f_N)$ . In renormalizable quantum field theories, fixing a finite number of parameters is sufficient to determine the  $n$ -point functions of a given finite set

of fields. In the case of the Weierstrass products, this is analogous to the situation where it is sufficient to take for the  $f_N$  polynomials of fixed order  $r$ . In this case, one can normalize  $P$  by demanding that  $P(0)$  and the first  $r$  derivatives of  $P$  at  $z = 0$  have prescribed values. This means that the solution  $P$  has  $r + 1$  free parameters. For  $r = 0$ , the solution is

$$P(z) = P(0) \lim_{N \rightarrow \infty} \prod_{i=1}^N (1 - z/z_i) .$$

For  $r = 1$  one obtains

$$P(z) = P(0) \exp(zP'(0)/P(0)) \lim_{N \rightarrow \infty} \prod_{i=1}^N (1 - z/z_i) \exp(z/z_i)$$

and so on.

When polynomials do not suffice, the number of free parameters becomes infinite. Quantum field theory is simpler, since the latter case does not seem to have an analogue. Moreover, quantum field theories are far more constrained than entire functions, since they only have a finite number of parameters, in contrast to the infinite set of the  $z_i$ .

For conformal field theories, the Weierstrass product formula is more than a far-fetched analogue, since many correlation functions involve Jacobi's theta-functions or Dedekind's  $\eta$ -function. Examples will be given below. Many important properties of these functions are best understood by their product formulas.

As one sees, regularization and renormalization are perfectly standard mathematical procedures. Their unfamiliar context was bound to cause some delay in understanding, but it is hard to comprehend how a delay of many decades could come about.

## 7 Structure Of Conformally Invariant Theories

One important way to deform a quantum field theory has not been introduced so far. One can change all  $n$ -point functions by a simple rescaling of the distances. When this change can be compensated by a transformation in  $L(F)$ , the theory is called scale invariant. More generally, the change is equivalent to such a transformation in addition to a change of the parameters of the theory. Infinitesimally, this equivalence is expressed by the Callan-Symanzik equation.

When a deformation should respect some symmetry, the corresponding field  $t(x)$  must be invariant under the symmetry group. In particular, this is true for Lorentz invariance. Indeed, our formalism does not require Lorentz invariance and can easily be adapted to quantum field theories on general spacetimes. One just has to replace the vector space  $F$  of fields by a bundle over spacetime. Let us conserve translational invariance, however, such that fields can be transported in canonical ways between arbitrary points of spacetime. When some component  $g_{\mu\nu}$  of the Riemannian metric is changed in a translationally invariant way, the corresponding field  $t$  is the component  $T^{\mu\nu}$  of the energy momentum tensor. For a rescaling of the distances, this yields  $t = T^\mu_\mu$ . For a scale invariant theory this means that the trace of the energy momentum tensor vanishes. Moreover, the integral  $\int T^{\mu\nu} d^D x$  must not depend on the distance scale, which means that the scaling dimension of the energy momentum tensor is equal to  $D$ .

Scale invariant quantum field theories are conformally invariant, too. This implies that the three point functions are known explicitly. The four-point functions reduce to functions of a single variable. Such theories have a good chance to be solvable in a rather explicit form, but for theories in more than two dimensions, the situation is still rather unclear. Nevertheless, recent developments indicate that these theories are important, too [Maldacena 1998, Witten 1998]. Suppose that you have a quantum field theory in  $k$  dimensional Minkowski space which admits a deformation to the



corresponding Anti-de-Sitter space. Recall that this is a homogeneous space of negative spatial curvature, with symmetry group  $SO(k-1, 2)$ . Anti-de-Sitter space has a  $(k-1)$ -dimensional boundary at infinity with a conformal structure, on which  $SO(k-1, 2)$  acts as the group of conformal transformations. When one takes suitable limits of the  $n$ -point functions, the theory in Anti-de-Sitter space reduces to a conformally invariant theory in a space of one lower dimension. In principle, the higher dimensional theory can be recovered from the boundary theory by techniques of algebraic quantum field theory [Rehren 1999].

Perhaps this procedure can be iterated. In this way, the properties of theories in higher dimension would be encoded in conformal field theories in two dimensions. This possibility is due to the typically quantum field theoretical fact that there is more freedom to construct conformal theories than higher dimensional quantum field theories in homogeneous spaces. In other words, the moduli spaces in higher spacetime dimensions have lower dimensions as manifolds, and can be embedded in the moduli spaces of quantum field theories in lower spacetime dimensions. As we shall see, string theory also performs such an encoding. It would be interesting to see if the two encodings are related.

In the following, we only will consider conformal field theories in two dimensions. The amount of technical details will just about suffice to put string theory in context. For a history of the crucial years 1984-88 and the relations to statistical mechanics, see [Itzykson, Saleur, Zuber 1988], which contains many references. A recent textbook is [di Francesco, Mathieu, Senechal 1997].

When one starts with a Minkowskian conformal field theory in flat spacetime, Wick rotation yields a euclidean theory on the Gauss plain. By conformal invariance, it is possible to compactify it to a theory on the Riemann sphere. As symmetry group, one obtains the group of linear rational transformations  $z \mapsto (az + b)/(cz + d)$  of the Riemann sphere. This will be the symmetry group of the  $n$ -point functions.

In two dimensions, the energy momentum tensor is a symmetric  $2 \times 2$  matrix. Because of scale invariance, its trace vanishes, such that it has only two independent components. By the Noether theorem, they are conserved quantities. More precisely, one linear combination is holomorphic, another anti-holomorphic. These are the famous Virasoro fields, which were first discovered in string theory [Virasoro 1970]. Their short-distance expansions are fixed by conformal invariance.

The symmetry transformations  $z \mapsto \lambda z$  introduce a change of the  $n$ -point functions which can be compensated by a linear transformation in  $L(F)$ . In most cases of interest, this transformation can be diagonalized. When a field transforms as  $\phi \mapsto \lambda^h \bar{\lambda}^{h'} \phi$ , we say that  $\phi$  has conformal dimensions  $(h, h')$ . When  $\lambda$  is real, we have a rescaling transformation. Thus  $h + h'$  is the scaling dimension of  $\phi$ . When  $|\lambda| = 1$ , we obtain a rotation, with an action described by the conformal spin  $h - h'$ . Since a rotation by  $2\pi$  is trivial, the conformal spin must be integral for bosonic fields. For holomorphic fields,  $h' = 0$ . Since the scaling dimension of the energy momentum tensor is 2, its holomorphic component has conformal dimensions  $(2, 0)$  and its anti-holomorphic component has conformal dimensions  $(0, 2)$ .

One could proceed in a purely algebraic way, completely within the framework for quantum field theories which was described above. Instead, let us shorten the path by some geometric intuition. Let us look at some holomorphic transformation  $z \mapsto f(z)$  of a neighborhood of  $z = 0$ . Locally, this is a symmetry, since it does not change the angles. When  $f(0) = 0$ , it induces a transformation in  $L(F)$ , since  $F$  can be considered as the space of fields at the point 0. The action of the transformations  $z \mapsto \lambda z$  on a field  $\phi$  of conformal dimensions  $(h, h')$  can be described by stating that the form  $\phi(z)(dz)^h(d\bar{z})^{h'}$  is invariant. If this remains true for all  $f$ , the field  $\phi$  is called primary. The primary fields span a subspace of  $F$ . If this subspace is finite dimensional, the corresponding conformal field theory is called minimal. The Ising model is minimal and has a three dimensional subspace of primary fields, but the Thirring model is not minimal.

The short distance expansion of a holomorphic field  $\phi$  of conformal dimensions  $(h, 0)$  on an arbitrary field  $\chi$  is a Laurent expansion, since it depends holomorphically on  $z$ . We write it in the form



$$\phi(z)\chi(w) = \sum_n (z-w)^{n-h} (\phi_n \chi)(w) .$$

For all integers  $n$ , this defines linear operators  $\phi_n$  on  $F$ . They are called the Fourier components of  $\phi$ . When  $\chi$  has conformal dimensions  $(\tilde{h}, \tilde{h}')$ , then  $\phi_n \chi$  has conformal dimensions  $(n + \tilde{h}, \tilde{h}')$ . We regard  $F$  as graded by the conformal dimensions and see that  $\phi_n$  is an operator of degree  $(n, 0)$ . The action of local conformal transformations on  $F$  is given by the linear operators  $L_n, \bar{L}_n$  obtained from the holomorphic and anti-holomorphic Virasoro fields.

For holomorphic fields  $\chi$ , the fields  $\phi_n \chi$  are holomorphic, too, such that one obtains a new algebraic structure [Zamolodchikov 1985, Borchers 1986, Goddard 1989]. A standard name in the physics literature is W-algebra, but mathematicians prefer to talk about vertex operator algebras. The latter name has the advantage of a clear history in string theory, whereas the W seems to be due to the accidental naming of some field as  $W(z)$  by Fateev and Zamolodchikov. Proposed allusions to Weyl, Wigner or Wilson are apocryphal, but may justify the name, which has the advantage of being short.

The field  $\phi_h \chi$  is called the normal ordered product of  $\phi$  and  $\chi$ . It is the first field which occurs in the regular part of the short distance expansion. In the Thirring model, the currents  $j, \bar{j}$  have conformal dimensions  $(1, 0)$  and  $(0, 1)$ . The Virasoro fields are given by the normal ordered products  $j_1 j / 2$  and  $\bar{j}_1 \bar{j} / 2$  [Callan, Dashen, Sharp 1967]. When  $n < h$ , the field  $\phi_n \chi$  occurs in the singular part. It turns out that it can be described in terms of commutators  $[\phi_n, \chi_m]$ . Thus one part of the operations of the W-algebra just describes a Lie algebra. For the components  $L_n$  of the energy momentum tensor this is the Virasoro algebra. It was discovered by Gelfand and Fuks [1968] and is a central extension of the Lie algebra of vector fields on a circle. The value of the central extension is universally called  $c$  for the holomorphic Virasoro field and  $\bar{c}$  for the anti-holomorphic one. In many models, they are equal. The values of  $c$  for the minimal models lie in a countable set. All of them have  $c < 1$ , whereas the Thirring model has  $c = 1$ . When  $\phi$  is holomorphic and  $\chi$  is anti-holomorphic, then  $[\phi_n, \chi_m] = 0$ .

The action of the  $\phi_n$  on the space  $F$  of all fields yields a representation of the W-algebra. With some effort, the representations of a fixed W-algebra can be given the structure of a tensor category, like the representations of a Lie algebra. The corresponding tensor product is called fusion product. The representation on the holomorphic fields themselves is called the basic representation and behaves as the neutral element under fusion.

Some W-algebras only have finitely many irreducible representations. These are called rational. In conformally invariant theories with rational W-algebras, all scaling dimensions are rational numbers. Such theories themselves are called rational, too. The minimal theories are characterized by the property that already the Virasoro part of the W-algebra has only finitely many irreducible representations. It is sufficient to consider the holomorphic Virasoro field, since for the anti-holomorphic one the situation is analogous. The properties of the Virasoro algebra only depend on the central extension  $c$ . The most interesting values occur for those minimal models where all the representations are unitary. This happens for  $c = 1 - 6/(p(p+1))$ ,  $p$  an integer greater 2. For  $p = 3$  one finds  $c = 1/2$  and the Ising model.

The first investigation of these questions was due to Mack and Lüscher. They found that  $c = 1/2$  is the lowest possible value and that there is a gap above  $1/2$ . Here is one of the rare cases where progress depended on difficult calculations performed by a mathematician. V. Kac determined the structure of the representations [1979], which later allowed Belavin, Polyakov and Zamolodchikov to determine the values of  $c$  for all minimal models [1984]. Soon afterwards, Friedan, Qiu and Shenker determined the unitary cases [1984].

The discovery of the minimal models and their explicit solution by Belavin, Polyakov, Zamolodchikov was the breakthrough event in the history of conformal field theory. It quickly became clear that these models are beautiful and fundamental mathematical structures. For reasons which are

hard to understand in depth, very different kinds of such structures, from Platonic solids to singularities, can be classified in terms of the ADE Dynkin diagrams. The same is true for the minimal models [Cappelli, Itzykson, Zuber 1987].

Part of the excitement about these early publications came from the relationship to continuous phase transitions in statistical mechanics. Besides the Ising model, many other well known continuous phase transitions were recognized as minimal models, like the ones for the 3-states Potts model, the tricritical Ising model, and the Lee-Yang edge singularity. Some of them have been realized in the lab, and measurements agree very well with the theoretical calculations.

Many properties of continuous phase transitions now fell into place. For example, some phase transitions are characterized by universal rational numbers, others have free continuous parameters. The former now are described by conformal field theories which have no conformally invariant deformations. In particular, this is true for the minimal models, like the continuum limit of the Ising model. When conformally invariant deformations exist, then they do not change  $c$ . The first example was Baxter's eight vertex model, which at the critical point becomes isomorphic to the older but more difficult Ashkin-Teller model. They yield  $c = 1$ , as for the closely related Thirring model. For a study of all unitary  $c = 1$  models, see [Ginsparg 1988].

In some respects even simpler than the minimal models are those for which the Virasoro fields can be described in terms of normal ordered products of fields of conformal dimensions  $(1,0)$  and  $(0,1)$ . Such fields are called currents, and the corresponding conserved integral quantities are called charges. The Thirring model is of this type, with currents  $j, \bar{j}$  and single holomorphic and anti-holomorphic charges  $j_0, \bar{j}_0$ . In more complex models where the two types of charges form simple Lie algebras, the short distance expansion of the currents yields the corresponding affine Kac-Moody algebras [Goddard, Olive 1988].

The Thirring model yields the simplest continuous family of conformal theories and has  $c = 1$ , too. In its bosonic description, it is given by the statistical mechanics of maps to a circle. The model is rational when the area of this circle is a rational number. This means that the set of points for which the model is rational is dense in the whole family. At one particular rational point, another continuous deformation is possible, which generates the moduli space of the Ashkin-Teller phase transitions. This is a first example of the rather intricate geometry of such moduli spaces, with many number theoretic aspects. As a first step, it would be important to know the rational points of more complex moduli spaces, since rational theories have very explicit descriptions. So far, there are very few results.

Within a moduli space of conformal theories, consider a perturbation by a field  $t(x)$ . The integral  $\int t(x) dz d\bar{z}$  must be invariant under conformal transformations, such that  $t$  should be a primary field of conformal dimensions  $(1,1)$ . In the Thirring model, the field  $j\bar{j}$  has these properties. The dimension of the vector space of such fields counts the number of possible infinitesimal deformations. Thus it is an upper bound on the dimension of the moduli space. For generic points of this space, one expects that the two dimensions are equal. For the Thirring model, they are both equal to 1.

The short distance expansion is a local property of the theory. When one wants to calculate the  $n$ -point functions, one also has to specify a Riemann surface on which the fields live (in the language of algebraic geometry, an algebraic curve). The simplest case is the Riemann sphere. Here the  $n$ -point functions of holomorphic fields are just rational functions. For more general fields, the results are much more complicated. For example, the four-point functions of minimal models already yield hypergeometric functions.

The Riemann sphere is unique, but more complicated Riemann surfaces (or equivalently algebraic curves) have their own continuous parameters. For example, a torus is described by the ratio  $\tau$  of two independent periods. When these are correctly ordered and varied continuously,  $\tau$  varies over the upper complex half-plane. The latter is called the Teichmüller space of the curves with torus topology. Points of Teichmüller space describe the same torus when they are related by a different choice of periods. Changes of the periods are described by the modular group. This is

the group of linear rational transformations  $\tau \rightarrow (a\tau + b)/(c\tau + d)$  with integral coefficients. More complicated curves behave in an analogous, but of course more complex way.

The 0-point function on the torus is essentially the partition function of the theory. Since energy and momentum are given by linear combinations of the Virasoro field components  $L_0$  and  $\bar{L}_0$ , the latter can be defined by

$$Z = \text{tr} \exp(2\pi i(L_0\tau - \bar{L}_0\bar{\tau})) ,$$

where the trace goes over the vector space  $F$  of all fields. The 0-point function on a torus with parameter  $\tau$  has the form

$$\tilde{Z} = \exp(-2\pi i(c\tau - \bar{c}\bar{\tau})/24) Z ,$$

where  $c, \bar{c}$  are the central extensions of the theory. The prefactor is necessary to get invariance under the modular group. For the Ising model one obtains

$$\tilde{Z} = \frac{1}{2} \sum_{i=2}^4 |\theta_i(\tau)/\eta(\tau)| ,$$

where the  $\theta_i$  are Jacobi's theta functions and  $\eta$  is Dedekind's function. Note that the scaling dimension  $1/8$  of the Ising spin can be read off from  $\theta_2$ .

For a free complex fermion one obtains

$$\tilde{Z} = \frac{1}{2} \sum_{i=2}^4 |\theta_i(\tau)/\eta(\tau)|^2 .$$

This function arises at the parameter  $R = \sqrt{2}$  of the Thirring model partition function

$$\tilde{Z} = |\eta(\tau)|^{-2} \sum_{m,n} \exp\left(\frac{\pi i}{2} \left(\left(\frac{m}{R} + nR\right)^2 \tau - \left(\frac{m}{R} - nR\right)^2 \bar{\tau}\right)\right) .$$

Here  $m, n$  vary over the integers. The equality of the latter two functions for  $R = \sqrt{2}$  is an example of the fermion-boson equivalence mentioned above. In the gaussian description, only the terms with  $m = n = 0$  were obvious, which explains why the model was considered to be uninteresting.

## 8 String Theory

Contrary to the historical developments, we have considered conformal field theory before coming to string theory. The reason is that string theory is more complex. Conformal field theory is just one ingredient, albeit an essential one. For general introductions to string theory and more references, see [Green, Schwarz, Witten 1987] and Polchinski [1998].

In 1968, Veneziano invented an amplitude for a scattering process with two incoming and two outgoing particles which shared several features with strong interaction processes. When a natural generalization to arbitrary particle numbers was found, Nambu, Nielsen and Susskind recognized that these amplitudes describe a one-dimensional object moving in space. The surface described by its motion is called a worldsheet. Its embedding into spacetime is described by functions  $X^\mu(\sigma, \tau)$ , where  $\sigma, \tau$  are coordinates on the worldsheet and  $X^\mu$  yields the corresponding spacetime positions.

Calculational problems arise, because there is no canonical parametrization of the worldsheet. Some natural choice can be made, however. The causal structure of the ambient spacetime induces a causal structure on the worldsheet, with two lightlike tangent directions at each point. These directions can be integrated to lightlike curves. One chooses the coordinates such that their equations are given by  $d\tau = d\sigma$  and  $d\tau = -d\sigma$ . This introduces a Minkowskian conformal structure on the  $\sigma, \tau$  parameter space. One chooses  $d\tau$  to be timelike and  $d\sigma$  to be spacelike.

Strings have finite spatial extent, such that the range of  $\sigma$  is compact. For open strings, the standard choice is an interval of length  $\pi$ , for closed strings a circle of circumference  $\pi$ . No further natural choices can be made, which means that the worldsheet dynamics is conformally invariant. In other words, the possible states of a single string are described by a conformal field theory. When one continues to a euclidean conformal field theory, one must make a Wick rotation in  $\tau$ , not in the time coordinate of  $X$ . The euclidean coordinate is called  $z$ .

By the analytic continuation, the worldsheet becomes a Riemann surface. Let us consider the case of closed strings only. Both in Minkowskian and in euclidean space, the worldsheet has the topology of a cylinder. By conformal invariance, it can be compactified to a Riemann sphere with two special points, one for the incoming and one for the outgoing state. Such special points are called punctures. String interactions are introduced by considering arbitrary Riemann surfaces with different numbers of punctures. Calculating the scattering of  $n$  strings involves three steps. First, the string states have to be identified with fields on the worldsheet. Secondly, one has to calculate the corresponding  $n$ -point functions for all Riemann surfaces with  $n$  punctures. The surfaces are not necessarily connected, since some groups of strings can interact independently of the others. Thirdly, one has to integrate over all of these configurations, in particular over the position of the punctures. In addition, one has to integrate over the finite dimensional moduli space of complex structures on Riemann surfaces with a given genus (number of handles). This integral is not needed when one applies conformal field theory to statistical systems, since there the Riemann surface is fixed.

Finally, one has to sum over the genera. Each term is multiplied by a power of the coupling constant. The exponent is proportional to an integral over the curvature, and can be normalized to  $g - 1$ . The leading contribution is given by  $g = 0$  and as many connected components as possible. For vanishing coupling, this leads to a free theory, exactly as for a quantum field theory. Indeed, a string theory can be regarded as a quantum field theory which includes graviton fields. In some limit, gravity decouples and one obtains a field theory of conventional type. For the latter, the perturbation series is a sum over Feynman diagrams. A tubular neighborhood of such a graph yields a Riemann surface of some genus  $g$ . This allows to identify one of the field theory couplings with the string coupling. We shall see that the others correspond to parameters of a conformal field theory on the string worldsheet.

The sum over  $g$  is certainly not convergent, which provides a technical reason to develop non-perturbative string theory. A deeper reason is the following. As for quantum field theory, free string theory can be considered as a boundary stratum on some moduli space. This stratum is characterized by the vanishing of a coupling constant, but in many cases its codimension is larger than one. Thus an expansion in the coupling constant cannot recover the full theory. In particular, it has no reason to be convergent. One example is given by quantum electrodynamics, where the following picture can be conjectured. To get a well defined quantum field theory, one has to introduce magnetic monopoles. These become infinitely heavy when the interaction goes to zero and their effects are not included in the perturbation expansion. Since monopoles can have an electric charge, one has an additional dimension of the moduli space which cannot be captured by perturbation theory.

In string theory, the rôle of the magnetic monopoles is taken over by branes of various dimensions. One can approach the full picture by a description of all possible boundary strata, but this goes much beyond the scope of the present article. Nevertheless, the reader should keep in mind that the following description of conformal worldsheet physics is perturbative and thus incomplete.

When a string state is described by a field  $\phi$  on the worldsheet, the integration over the corresponding puncture position takes the form  $\int \phi(z) dz d\bar{z}$ . This must make sense independently of the choice of the coordinate  $z$ . In other words, string states are described by primary fields of conformal dimensions (1,1). There is another way to get the same result. When the string is considered in the background of some particle wave in spacetime, this yields a conformally invariant deformation of the theory, at least infinitesimally. Since deformations are described by the primary fields of conformal dimensions (1,1), the same must be true for the particle states arising from the

string. With reference to spontaneous symmetry breaking in quantum field theory, the existence of particle states may be described as a Goldstone phenomenon.

When one considers strings in flat spacetime, the coordinates of the latter can be regarded separately. For a space coordinate  $X^i$  appropriate fields are given by  $\exp(ip_i X^i(z))$ , with arbitrary  $p_i$ . With a conventional choice of the length scale, the scaling dimension of this field is  $p_i^2/4$ .

A new situation appears for the time coordinate  $X^0$ . Due to Lorentz invariance, the field  $\exp(ip_0 X^0(z))$  has scaling dimension  $-p_0^2/4$ . Fields with negative scaling dimensions of arbitrary size do not occur in statistical mechanics, but they can be made to fit in the framework of conformal field theory. Indeed, without such negative contributions to the scaling dimension, one never would get an infinite number of particle states. Here we can take an arbitrary field with  $h = h'$  and adjust the value of  $p_0^2$  such that the scaling dimension becomes 2. This produces at least a (1,1) field, though in general it will not be primary.

When we disregard the latter problem, we can consider the fields  $\partial X^\mu \bar{\partial} X^\nu \exp(ipX)$ . They have conformal dimensions (1,1) when  $p^2 = 0$ , such that they describe massless particles. When one considers their behaviour under spacetime rotations, one sees that they include spin 2 particles, i.e. states which behave like gravitons. For general reasons, the coupling of such states must be described by Einstein's theory. Thus any consistent string theory is a theory of quantum gravity.

Later, this fact was recognized as the best feature of string theory, but it was a nuisance as long as the theory was supposed to work for the strong interaction. Other problems of the original string theory had to be solved quite apart of this deeper issue, namely the existence of tachyons and the wrong dimension of spacetime.

A tachyon appears when one considers the simple field  $\exp(ipX)$ . This is a primary (1,1) field, if  $p^2 = -8$ . To get rid of this unwanted particle with negative squared mass, the conformal symmetry had to be extended to a superconformal one. The fields of such theories can have integral or half-integral conformal spin. Those with a half-integral difference  $h - h'$  are fermionic. In addition to the Virasoro fields one has fields  $G$  and  $\bar{G}$  of conformal dimensions (3/2,0) and (0,3/2). There are two different fermion numbers associated to the holomorphic and anti-holomorphic variables. The short distance expansion with  $G$  changes the first one by one unit, that with  $\bar{G}$  the second one. The Fourier components  $L_n$  of the Virasoro field and those of  $G$  together yield a superalgebra, in which the Virasoro algebra is embedded. The model has two sectors (discovered separately by Ramond and by Neveu and Schwarz), but we shall consider just the latter. In this sector, the fermionic fields have half-integral coefficients. Apart from such modifications, superconformal field theory can be regarded as a special case of conformal field theory, so most of the preceding description remains valid.

Fields related by the action of  $G_{1/2}, \bar{G}_{1/2}$  are called superpartners. For superconformal deformations, the corresponding (1,1) fields must be superpartners of (1/2,1/2) fields. The physically relevant deformations are described by bosonic fields, such that the (1/2,1/2) fields must be fermionic with respect to both fermion numbers. The superstring still has fields  $\exp(ipX)$ , which have conformal dimensions (1/2,1/2) for  $p^2 = -4$ , but these fields are of bosonic nature and do not correspond to physical particles. This elimination of the tachyonic fields is due to Gliozzi, Olive and Scherk.

The issue of the spacetime dimension arose in a different way. When one calculates the norm of a field of type  $\partial X^\mu$ , Lorentz invariance yields a result proportional to  $g_{\mu\mu}$ . In particular, one can find negative norms which are incompatible with a probability interpretation. In the 50's and 60's much ink had flown in unsuccessful attempts to make sense out of negative norms and no one was motivated to try again. Fortunately, Virasoro recognized that not all fields yield physical states [1970]. The concepts of primary fields and conformal dimensions did not exist yet, but he only found the correct constraints and described them by the Fourier modes of the Virasoro fields. One year later, Galli obtained the interpretation in terms of conformal invariance [1970].

Numerical investigations showed up to a certain degree of complexity that the physically allowed fields all have positive norm, but a general proof was difficult to obtain. Then it turned out that

allowed negative norm fields do exist when the spacetime dimension is greater than 26, or 10 for the superstring. This made sense of an observation of Lovelace [1971], which had not been taken very seriously because it was too outlandish. Looking at Riemann surfaces of torus topology, Lovelace had shown that the bosonic string theory was found to require a spacetime of 26 dimensions. Now it became clear that this number was a deep structural property of the bosonic string theory and would not go away. Indeed, Brower [1972] and Goddard and Thorn [1972] used the 26 dimensions to prove that the norms make physical sense (the no-ghost theorem). Later it turned out that the value of this critical dimension has deep relations to the conformal invariance of the world sheet physics and the corresponding modular invariance [Brink, Nielsen 1973]. Moreover, Beilinson and Manin found out that the strange 26 was closely related to analytic torsion results of Mumford, which allowed them to write the measure for the integration over the moduli space of Riemann surfaces in a very elegant form [1986].

The critical dimension translates into the value  $c = \bar{c} = 26$  of the central extensions. For the superstring one needs 10 dimensions and  $c = \bar{c} = 15$ . The latter value is due to the superpartners of the 10 coordinates  $X^\mu$ , which contribute half as much to the central extension. The simplest way to obtain a model in four dimensions is the old Kaluza-Klein idea. One just wraps up all superfluous dimensions in a small circle. For the bosonic string this yields 22 copies of the Thirring model. The corresponding 44 currents of type  $j$  and  $\bar{j}$  yield 44 photons, all with separate interactions of electromagnetic type. The values of  $c, \bar{c}$  do not change. Obviously, this model is not particularly realistic. It exemplifies, however, that the spacetime dimension of the model can be changed at will, as long as one keeps conformal invariance and the correct central extensions. For the superstring, similar remarks apply.

To write down a general bosonic string model in four dimensions, one just needs to replace the 22 copies of the Thirring model by an arbitrary conformal field theory with  $c = \bar{c} = 22$ . The latter is called the internal conformal theory. In analogy to the Kaluza-Klein case, one still says that it describes 22 compactified dimensions, even if this is not always a geometrically correct interpretation. To compactify the superstring to fourdimensional spacetime, one needs six compactified dimensions and  $c = \bar{c} = 9$  instead. Every possible compactification corresponds to a theory in a space of less than 10 spacetime dimensions. In this way one gets, e.g., an encoding of four-dimensional quantum field theories by conformal or superconformal field theories in two dimensions.

In particular, consider a field  $\phi(z) \exp(iXp)$ , where  $\phi$  belongs to the internal theory. When one adjusts  $p^2$  to get overall conformal dimensions (1,1), one sees that for a particle state of mass  $m$  the corresponding field must have contributions  $h = h' = 1 + m^2/8$  from the internal conformal theory. Of particular interest are the internal (1,1) fields, which correspond to massless Higgs bosons.

When the conformal theory includes an affine Kac-Moody algebra with holomorphic currents  $j_a$  of conformal dimensions (1,0), the fields  $j_a \bar{\partial} X^\mu \exp(iXp)$  with  $p^2 = 0$  describe the quanta of a vector potential  $A_a^\mu$  belonging to the corresponding finite dimensional gauge group. Thus the states of the string now include non-abelian gauge fields, and the theory starts to look a bit more like the standard model. In the superstring theory, one also gets fermions. Their interactions with the Higgs bosons and the gauge fields are of standard type, though one has not yet managed to obtain precisely the standard model.

Of course, the bosonic model always will have the tachyonic  $\exp(ipX)$  fields and cannot be used by itself. Nevertheless, the bosonic string can be used for either the holomorphic or the anti-holomorphic coordinates. To get rid of the tachyon, it is indeed sufficient to use a field  $G$  but no  $\bar{G}$ . This yields models with  $c = 26$  but  $\bar{c} = 15$ , called heterotic string models. They were found by Gross, Harvey, Martinec and Rohm. Heterotic strings do not have a pure spacetime version, since the spacetime contributions to  $c$  and  $\bar{c}$  have to match. The archetypal heterotic string lives in 10 dimensions, where the compactified part is purely holomorphic, with  $c = 16$ . There are many constraints on purely holomorphic conformal theories which exist on arbitrary Riemann surfaces. In particular,  $c$  must be a multiple of 8. For  $c = 8$ , the only example is the affine Kac-Moody algebra based on  $E_8$ . For  $c = 16$ , one either can take the tensor product of two  $c = 8$  models or use

the affine Kac-Moody algebra based on  $SO(32)$ . For  $c = 24$ , there are 71 possibilities [Schellekens 1993]. One of them has a remarkable symmetry group of about  $10^{54}$  elements, the Fischer-Griess monster [Borcherds 1986]. This closeness of string theory to beautiful exotic structures is still a deep mystery. To get to four dimensions, one needs a less exotic internal conformal field theory with  $c = 22$  and  $\bar{c} = 9$ .

The late discovery of the heterotic string was due to the fact that string research slowed down a lot after 1974. At that time it had become clear that QCD is a better theory for the strong interaction. Though Scherk and Schwarz had shown that one could reinterpret string theory as a theory of quantum gravity [1974], there was not very much support for such an arcane research line. One of the ideas which appeared shortly before the theory entered a long hibernation period was the classification of the possible rational theories by modular forms. In particular, this yielded a candidate which later would be interpreted as the partition function of the compactified part of the heterotic string [Nahm 1977]. Unfortunately, present mathematical techniques only allow to apply this procedure to rational conformal fields theories, for which the partition function can be written as a finite sum  $Z = \sum_i Z_i \bar{Z}_i$  with holomorphic functions  $Z_i$  and anti-holomorphic functions  $\bar{Z}_i$ . Nevertheless, it was striking that the method indicated an incredibly large number of possible theories, in stark contrast to the initial hopes that one was heading for something unique. At present, the situation has not changed very much. There are vague hopes that non-perturbative string theory will select particular models, but it also is possible that one will end up on a moduli space with more parameters than in the standard model. For every choice of parameters, one will have a quantum version of Einstein's gravity theory, however.

The bold switch from the interpretation of string theory as a theory of the strong interaction to a theory of quantum gravity by Scherk and Schwarz must have been one of the strangest events in the history of physics. In particular, the basic distance scale had to be changed by twenty orders of magnitude, from the proton diameter to the Planck length. For mathematicians this should be easier to digest than for physicists, since no change in the mathematical structure is involved. For physicists, however, the emergence of string theory now appears as an accident. It would not have happened if the discovery of the  $SU(3)$  gauge interaction of the quarks had come a few years earlier. Even in hindsight, one sees no way how a direct study of quantum gravity could have led to this theory. Indeed, a direct attack has been tried from several points of view, but with very limited success. It seems that one cannot unify quantum theory and gravitation without incorporating much knowledge about other interactions.

In any case, present research on quantum gravity cannot follow the traditional pattern of physics. One hundred years ago, Planck himself estimated its characteristic length scale by combining Newton's constant, the speed of light, and his new quantum of action. He found  $4,13 \cdot 10^{-33} \text{ cm}$  [1900]. At that time, physicists and chemists were getting the first precise ideas about what happens at  $10^{-8} \text{ cm}$ , so Planck must have felt like looking into an abyss. Of the 25 orders of magnitude to be covered we now have explored not quite ten, thus a naive extrapolation predicts another 150 years before we really understand what is going on at the basic scale. Planck's report about his discovery is brief and sober. Nevertheless, he states that the units he found would keep their meaning for all times and all cultures, including extraterrestrial and non-human ones.

Without the ability to do experiments in quantum gravity, it is hard to know if theoretical investigations are on the right track. Everyone who keeps trying is inspired by Einstein's success with general relativity. He had little experimental input, but relied on his keen sense for structure and mathematical beauty. His belief in the harmony of the spheres was as deep as Kepler's, and when he had found an indication of congruence between nature and a mathematical structure he did everything to uncover it fully. Quantum mechanics remained as a jarring note like the irrational numbers to the early Pythagoreans. Thus string theory would have disappointed Einstein as far as quantum physics is concerned. But if it is correct, it justifies some of his attempts in the search for a unified theory. On one hand, he wanted to generalize the metric tensor  $g_{\mu\nu}$  to an object with an antisymmetric part. String theory has such an object, called the  $B$  field. Together with the metric tensor it is obtained from the fields  $\partial X^\mu \bar{\partial} X^\nu \exp(ipX)$  which have been considered above.



Einstein also was right in his high regard for the Kaluza-Klein approach.

Einstein's example can be used as an encouragement and as a warning. His successful gravity theory was based on at least one elementary fact which no one else could explain - the equality of inertial and gravitational mass. When he let loose of such guidance, he still did important research, but went astray. String theory does not do too badly on this account. On one hand, one can at least come close to the standard model. Moreover, superstring theory at least suggests that the experimentalists will find supersymmetry in the near future. The theory develops in a search for deep and beautiful structures, but it has the advantage of holding on to the tenuous guide offered by low energy experiments. Currently, no other theory of quantum gravity can make such claims. Despite its unbelievable origin, string theory is by far the most promising approach to unify all of the known interactions. The one possible exception of the latter claim is the cosmological constant, since it is separated by another abyss of many orders of magnitude from the rest of physics.

## 9 Missed And Open Opportunities

Let us come back to Dyson's 1972 address to the American Mathematical Society. It was titled 'Missed Opportunities' and concerned problems in the communication between mathematicians and physicists. In particular, he considered quantum field theory and the unification with gravity, but his first example concerned a communication problem between Dyson the physicist and Dyson the mathematician. As a mathematician, he had played around with powers of Dedekind's  $\eta$ -function and obtained nice identities for the exponents 3,8,10,14,15,21,24,26,28,35,36,... In this combinatorial context he did not recognize the dimensions of the simple Lie groups, which would have been evident to him in a physics context. From today's point of view, the regret about this little failure may have caused him to miss a greater opportunity. He must have heard about the incredible 26 dimensions of the bosonic string, which Lovelace had found the year before, but apparently thought little about this coincidence. Otherwise he would have stumbled on the importance of Dedekind's  $\eta$ -function in string theory. Two years later Scherk and Schwarz established the importance of string theory for the unification of quantum field theory and gravity.

Many of the present author's missed and taken opportunities also concern the interaction with mathematics. A very rapid course by D. Zagier led to a classification of string theories by modular functions. On the other hand, searching the CERN library for books discussing infinite dimensional Lie algebras was a frustrating enterprise. Even worse was the inability to find the dimensions of the next representations of  $E_8$ , when the classification yielded  $q^{-1/3} + 248q^{2/3} + \dots$ . Unfortunately, the visit of Kac to CERN came far too late, but to the author it proved the value of an environment where physicists and mathematicians could make the effort to learn about their respective discoveries. Princeton and some other places made a good start, but it would be nice to have a few more. Here are some problems which may be tackled in such an environment.

In the Kaluza-Klein formalism, a fifth dimension is hiding because it is compactified to a circle. When one considers experiments at fixed energy and makes the period of the compactified dimension very large, the five dimensional geometry emerges again. This process can be generalized - taking suitable limits of quantum field theories one can obtain classical geometries. In this context, the latter are called target spaces.

Let us look at the Kaluza-Klein situation in the context of string theory. We have considered fields  $\exp(iXp)$ , where for simplicity we consider a single position component  $X$ . When it is compactified with period  $l$ , the choice of  $p$  is constrained by  $\exp(ilp) = 1$ , or  $p = 2\pi n/l$  with integral  $n$ . The scaling dimension of such a field is proportional to  $(n/l)^2$ , thus small for large  $l$  and small  $n$ . In particular, the integer  $h - h'$  has to vanish. The short distance expansion of such fields involves weak singularities only. In the limit where  $l$  becomes large, it reduces to the ordinary product  $\exp(iXp_1) \exp(iXp_2) = \exp(iX(p_1 + p_2))$ . Because of scale invariance, the large  $l$  limit produces a unique commutative algebra of all fields whose scaling dimension approaches zero.



In the classical limit, every smooth function on the Kaluza-Klein circle can be Fourier expanded with basis  $\exp(iXp)$ . Thus one obtains the algebra of all smooth functions on the circle. Moreover, the space of these functions is graded by the scaling dimension  $2h$ , which is the eigenvalue of the Laplace operator. The geometry of the circle can easily be reconstructed from this information.

This example can be generalized to all kinds of manifolds. Particularly attractive are Calabi-Yau manifolds, where one can work with the highly constrained superconformal theories. Much less is known about these manifolds than about the circle, for example about their Einstein metrics. In these cases, the conformal theories may be easier to control than their classical limits. One certainly may hope to obtain the algebra of smooth functions and the corresponding eigenvalues of the Laplace operator from the conformal data.

When the parameters of a conformal field theories are varied, one may obtain quite different classical limits. In particular, a connected moduli space may have several different boundary components. In this way, it is possible to relate different classical geometries by non-classical paths. As a simple example, consider again a string on the Kaluza-Klein circle. There are more fields than we have considered so far, since one can wind the string around the circle. When the circle is large, this yields particle states of large mass. When the period  $l$  becomes very small, winding costs hardly any energy, whereas the  $\exp(iXp)$  fields have large scaling dimensions and describe particles of large mass. When  $l$  goes to zero, the short distance expansion of the basic winding fields is just given by the additive group of the winding numbers. In this way, a Kaluza-Klein theory with period  $l$  becomes isomorphic to one with period  $l^{-1}$ . This isomorphism is known as T-duality. It is one of many dualities which arise in string theory, so the name 'dual model' used around 1970 was quite prescient. Perhaps the most famous of the dualities is mirror symmetry, which is a specific property of conformal field theories with a high degree of supersymmetry.

The winding fields are examples of solitonic objects, since the winding number is time-independent. In the euclidean conformal theory, one also has instanton contributions given by a map of the string Riemann surface to the target space. The most studied case is the one of embeddings of Riemann spheres in algebraic target manifolds, since mathematicians have been much interested in another abyss of many orders of magnitude from the rest of physics.

### Missed and open opportunities

Let us come back to Dyson's 1972 address to the American Mathematical Society. It was titled 'Missed Opportunities' and concerned problems in the communication between mathematicians and physicists. In particular, he considered quantum field theory and the unification with gravity, but his first example concerned a communication problem between Dyson the physicist and Dyson the mathematician. As a mathematician, he had played around with powers of Dedekind's  $\eta$ -function and obtained nice identities for the exponents 3,8,10,14,15,21,24,26,28,35,36,... In this combinatorial context he did not recognize the dimensions of the simple Lie groups, plus 26, which would have been evident to him in a physics context. From today's point of view, the regret about this little failure may have caused him to miss a greater opportunity. He must have heard about the incredible 26 dimensions of the bosonic string, which Lovelace had found the year before, but apparently thought little about this coincidence. Otherwise he would have stumbled on the importance of Dedekind's  $\eta$ -function in string theory. Two years later Scherk and Schwarz established the importance of string theory for the unification of quantum field theory and gravity.

Many of the present author's missed and taken opportunities also concern the interaction with mathematics. A very rapid course by D. Zagier led to a classification of string theories by modular functions. On the other hand, searching the CERN library for books discussing infinite dimensional Lie algebras was a frustrating enterprise. Even worse was the inability to find the dimensions of the next representations of  $E_8$ , when the classification yielded  $q^{-1/3} + 248q^{2/3} + \dots$ . Unfortunately, the visit of Kac to CERN came far too late, but to the author it proved the value of an environment where physicists and mathematicians could make the effort to learn about their

respective discoveries. Princeton and some other places made a good start, but it would be nice to have a few more. Here are some problems which may be tackled in such an environment.

In the Kaluza-Klein formalism, a fifth dimension is hiding because it is compactified to a circle. When one considers experiments at fixed energy and makes the period of the compactified dimension very large, the five dimensional geometry emerges again. This process can be generalized – taking suitable limits of quantum field theories one can obtain classical geometries. In this context, the latter are called target spaces.

Let us look at the Kaluza-Klein situation in the context of string theory. We have considered fields  $\exp(iXp)$ , where for simplicity we consider a single position component  $X$ . When it is compactified with period  $l$ , the choice of  $p$  is constrained by  $\exp(ilp) = 1$ , or  $p = 2\pi n/l$  with integral  $n$ . The scaling dimension of such a field is proportional to  $(n/l)^2$ , thus small for large  $l$  and small  $n$ . In particular, the integer  $h - h'$  has to vanish. The short distance expansion of such fields involves weak singularities only. In the limit where  $l$  becomes large, it reduces to the ordinary product  $\exp(iXp_1)\exp(iXp_2) = \exp(iX(p_1 + p_2))$ . Because of scale invariance, the large  $l$  limit produces a unique commutative algebra of all fields whose scaling dimension approaches zero.

In the classical limit, every smooth function on the Kaluza-Klein circle can be Fourier expanded with basis  $\exp(iXp)$ . Thus one obtains the algebra of all smooth functions on the circle. Moreover, the space of these functions is graded by the scaling dimension  $2h$ , which is the eigenvalue of the Laplace operator. The geometry of the circle can easily be reconstructed from this information.

This example can be generalized to all kinds of manifolds. Particularly attractive are Calabi-Yau manifolds, where one can work with the highly constrained superconformal theories. Much less is known about these manifolds than about the circle, for example about their Einstein metrics. In these cases, the conformal theories may be easier to control than their classical limits. One certainly may hope to obtain the algebra of smooth functions and the corresponding eigenvalues of the Laplace operator from the conformal data.

When the parameters of a conformal field theories are varied, one may obtain quite different classical limits. In particular, a connected moduli space may have several different boundary components. In this way, it is possible to relate different classical geometries by non-classical paths. As a simple example, consider again a string on the Kaluza-Klein circle. There are more fields than we have considered so far, since one can wind the string around the circle. When the circle is large, this yields particle states of large mass. When the period  $l$  becomes very small, winding costs hardly any energy, whereas the  $\exp(iXp)$  fields have large scaling dimensions and describe particles of large mass. When  $l$  goes to zero, the short distance expansion of the basic winding fields is just given by the additive group of the winding numbers. In this way, a Kaluza-Klein theory with period  $l$  becomes isomorphic to one with period  $l^{-1}$ . This isomorphism is known as T-duality. It is one of many dualities which arise in string theory, so the name 'dual model' used around 1970 was quite prescient. Perhaps the most famous of the dualities is mirror symmetry, which is a specific property of conformal field theories with a high degree of supersymmetry.

The winding fields are examples of solitonic objects, since the winding number is time-independent. In the euclidean conformal theory, one also has instanton contributions given by a map of the string Riemann surface to the target space. The most studied case is the one of embeddings of Riemann spheres in algebraic target manifolds, since mathematicians have been much interested in counting the number of different embeddings. As shown by Candelas' group, mirror symmetry yields the correct numbers for the quintic in four dimensional projective space. This started the huge interest of mathematicians in this topic. Usually, mathematicians try to replace the quantum field theoretic approach by more classical methods, but in the end this may well turn out to be the more arduous approach, quite comparable to a proof of the prime number theorem without using analysis.

The moduli space of superconformal theories of fixed central extensions seems to be connected. For  $c = \bar{c} = 9$  each Calabi-Yau manifold of three complex dimensions is a possible target space and

yields one boundary component of the moduli space. By following all ramifications of the moduli space it should be possible to classify all such Calabi-Yau manifolds, for a start.

In modern algebraic geometry, geometric and number theoretic problems occur side by side. The same is true of conformal field theory, though physicists so far are ill equipped to handle these issues. For example, what is the meaning of Dyson's formula for  $\eta^{26}$ ? For the moment, this is a mystery without much of a clue, but one can start with simpler problems. Above, we have discussed rational points in the moduli space of string theories. At these points, coupling constants like the Yukawa couplings can be calculated explicitly. Usually, the rational models are among the few which such calculations are possible at present. For example, let us take models with  $c = \bar{c} = 6$  and sufficiently large supersymmetry. In this case one obtains K3 surfaces as target space. The moduli space turns out to have 80 dimensions, but the largest submanifold under explicit control has just 16 dimensions. In addition, however, there are many rational points sprinkled around which are perfectly well understood. Will the rational points turn out to be densely distributed? More specifically, the same moduli space occurs for torus compactifications of the heterotic string to six-dimensional flat spacetime. In the latter case the rational points are well known, and they correspond precisely to the complex multiplication points of the K3 moduli space. Are those the rational points of the latter moduli space?

In simple examples, the conformal dimensions of rational models are obtained by applying the dilogarithm function to algebraic numbers. The corresponding sums are described by torsion elements in the Bloch group [Nahm, Recknagel, Terhoeven 1993]. Thus there is a link between conformal field theory and one of the most active areas of present mathematical research. In particular, there seem to be relations to Grothendieck's program for a description of the Galois group of all algebraic numbers and to the theory of motives. Kontsevich recently conjectured that the motivic Galois group acts on the moduli space of conformal field theories [1999].

Obviously, mathematicians have much to gain from physics. In view of the higher reliability of the answer (and regarding costs as irrelevant) physicists were more inclined to ask nature than to ask a mathematician. Quite typically, Schweber's textbook concludes with the following sentences: "In the final analysis, however, it will probably be the new information that will be obtained from the high energy machines and colliding beam machines to go into operation in the next few years which will help unravel the puzzle of the elementary particles and their interactions. In particular, we may discover whether the notions of space and time upon which present-day field theories are based are in fact valid." But meanwhile we have learned more respect for the sixteen orders of magnitude which separate us from the Planck scale. If it gets too expensive to ask the direct questions, we just have to push the mathematical analysis of what little clues there are. There is hope, since sometimes it did work. Kepler managed to extract the secrets of the planetary motions from pretelescopic data, but it would have been much harder without some knowledge about ellipses.

As a final encouragement for those willing to use the bridge, let me quote a German poet: "Nur Beharrung führt zum Ziele, nur die Fülle führt zur Klarheit und im Abgrund wohnt die Wahrheit" (to reach the goal you must be persistent, to see clearly you have to understand a wealth of phenomena, and truth lives in the abyss). Schiller's poem talks about causal time and three-dimensional space, but two euclidean dimensions make a good start.

## References

- [1] M.F. Atiyah, N.J. Hitchin, I.M. Singer, Self-duality in four-dimensional Riemannian geometry, *Proc. Roy. Soc. L. A* 362, p. 425, 1978
- [2] H. Bateman, The transformation of the electrodynamic equations, *Proc. London Math. Soc.* 8, p. 223, 1910
- [3] E. Bessel-Hagen, Über die Erhaltungssätze der Elektrodynamik, *Math. Ann.* 84, p. 258, 1921

- [4] A.A. Beilinson, Yu.I. Manin, The Mumford form and the Polyakov measure in string theory, *Comm.Math. Phys.* 107, p. 359, 1986
- [5] A.A. Belavin, A.M. Polyakov, A.B. Zamolodchikov, Infinite conformal symmetry in two-dimensional quantum field theories, *Nucl.Phys. B* 241, p. 333, 1984
- [6] L. Brink, H.B. Nielsen, A physical interpretation of the Jacobi imaginary transformation and the critical dimension in dual models, *Phys.Lett.* 43B, p. 319, 1973
- [7] R.C. Brower, Spectrum-generating algebra and no-ghost theorem for the dual model, *Phys.Rev. D* 6, p. 1655, 1972
- [8] N. Bohr, L. Rosenfeld, Zur Frage der Messbarkeit der elektromagnetischen Feldgrößen, *Dan.Math.Fys.Medd.* 12, Nr. 8, 1933
- [9] R.E. Borcherds, Vertex algebras, Kac-Moody algebras and the monster, *Proc.Nat.Acad.Sci. USA* 83, p. 3068, 1986
- [10] C.G. Callan, S. Coleman, R. Jackiw, A new improved energy-momentum tensor, *Ann.Phys.* 59, p. 42, 1970
- [11] C.G. Callan, R.F. Dashen, D.H. Sharp, Solvable two-dimensional field theory based on currents, *Phys. Rev.* 165, p. 1883, 1968
- [12] A. Cappelli, C. Itzykson, J.-B. Zuber, The A-D-E classification of minimal and  $A_1^1$  conformal invariant theories, *Comm.Math.Phys.* 113, p. 1, 1987
- [13] S. Coleman, Quantum sine-Gordon equation as the massive Thirring model. *Phys.Rev. D* 11, p. 2088, 1975
- [14] R.P. Crease and C.C. Mann, *The Second Creation*, Macmillan, New York 1987
- [15] E. Cunningham, The principle of relativity in electrodynamics and an extension thereof, *Proc. London Math. Soc.* 8, p. 77, 1910
- [16] P. di Francesco, P. Mathieu, D. Senechal, *Conformal Field Theory*, New York 1997
- [17] H.P. Dürr, W. Heisenberg, H. Mitter, S. Schlieder, K. Yamazaki, Zur Theorie der Elementarteilchen, *Zeits. Naturfor.* 14a, p. 441, 1959
- [18] F.J. Dyson, The radiation theories of Tomonaga, Schwinger, and Feynman, *Phys.Rev.* 75, p. 486, 1949
- [19] F.J. Dyson, Missed opportunities, *Bull. AMS*, p. 635, 1972
- [20] S. Ferrara, R. Gatto, A.F. Grillo, Conformal algebra in two space-time dimensions and the Thirring model, *Nuovo Cim.* 12A, p. 959, 1972
- [21] D. Friedan, Z. Qiu, S. Shenker, Conformal invariance, unitarity and critical exponents in two dimensions, *Phys.Rev.Lett.* 52, p. 1575, 1984
- [22] A. Galli, Conformal invariance in the dual symmetric theory of hadrons, *Nuovo Cim.* 69A, p. 275, 1970
- [23] I.M. Gelfand, D.B. Fuks, Cohomologies of the Lie algebra of the vector fields on the circle, *Funct.Anal.Appl.* 2, p. 342, 1968
- [24] M. Gell-Mann, Symmetries of baryons and mesons, *Phys. Rev.* 125, p. 1067, 1962
- [25] P. Ginsparg, Curiosities at  $c=1$ , *Nucl.Phys. B* 295, p. 153, 1988

- [26] P. Goddard, Meromorphic conformal field theory, in: Infinite Dimensional Lie Algebras and Lie Groups, V.G. Kac ed., Adv.Ser.Math.Phys. 7, p. 556, World Scientific, 1989
- [27] P. Goddard, C.B. Thorn, Compatibility of the dual Pomeron with unitarity and the absence of ghosts in the dual resonance model, Phys.Lett. 40B, p. 235, 1972
- [28] P. Goddard, D. Olive, eds., Kac-Moody and Virasoro Algebras, World Scientific, Singapore 1988
- [29] M. Green, J. Schwarz, E. Witten, Superstring Theory, Cambridge University Press 1987
- [30] F. Gürsey, On a conform-invariant spinor wave equation, Nuovo Cim. 3, p. 988, 1956
- [31] R. Haag, On quantum field theory, Dan.Math.Fys.Medd. 20, p. 12, 1955
- [32] E.M. Henley, W. Thirring, Elementary Quantum Field Theory, McGraw-Hill, New York 1962
- [33] W.V.D. Hodge, The Theory and Application of Harmonic Integrals, Cambridge Univ. Press 1941
- [34] C. Itzykson, H. Saleur, J.-B. Zuber, eds., Conformal Invariance and Applications to Statistical Mechanics, World Scientific, Singapore 1988
- [35] K. Johnson, Solution of the equation for the Green's functions of a two dimensional relativistic field theory, Nuovo Cim. 20, p. 773, 1961
- [36] V.G. Kac, Contravariant form for infinite dimensional Lie algebras and superalgebras, Lecture Notes in Physics 94, p. 441, Springer 1979
- [37] L.P. Kadanoff, Scaling laws for Ising models near  $T_c$ , Physics 2, p. 263, 1966
- [38] L.P. Kadanoff, Operator algebra and the determination of critical indices, Phys.Rev.Lett. 23, p. 1430, 1969
- [39] O. Klein, On the theory of charged fields, in: New Theories in Physics, Warsaw 1938, Proc., Nyhoff, The Hague 1939
- [40] M. Kontsevich, Operads and motives in deformation quantization, math.QA/9904055
- [41] J. Lepowsky, R.L. Wilson, Construction of the affine Lie algebra  $A_1^{(1)}$ , Comm.math.Phys. 62, p. 43, 1978
- [42] C. Lovelace, Pomeron form factors and dual Regge cuts, Phys.Lett. 34B, p. 500, 1971
- [43] J. Maldacena, The large N limit of superconformal field theories and supergravity, Adv.Theor.Math.Phys. 2, p. 231, 1998
- [44] T. Matsubara, A new approach to quantum-statistical mechanics, Prog.Theor.Phys. 14, p. 351, 1955
- [45] W. Nahm, Spin in the spectrum of states of dual models, Nucl.Phys. B120, p. 125, 1977
- [46] W. Nahm, A. Recknagel, M. Terhoeven, Dilogarithm identities in conformal field theory, Mod.Phys.Lett. 18, p. 1835, 1993
- [47] Y.J. Ng; Julian Schwinger, the Physicist, the Teacher, and the Man, World Scientific, Singapore 1996
- [48] E. Noether, Invarianten beliebiger Differentialausdrücke, Nachr. d. Göttinger Akad. d. Wiss. 1918, p. 235

- [49] L. Onsager, A two-dimensional model with an order-disorder transition, *Phys.Rev.* 65, p. 117, 1944
- [50] K. Osterwalder, Euclidean Green's functions and Wightman distributions, in: *Constructive quantum field theory*, Lecture Notes in Physics 25, Springer 1973
- [51] A. Pais, *Inward Bound*, Oxford University Press, 1986
- [52] W. Pauli, Über die Invarianz der Dirac'schen Wellengleichungen gegenüber Ähnlichkeitstransformationen des Linienelementes im Fall verschwindender Ruhmasse, *Helv.Phys.Acta* 13, p. 204, 1940
- [53] M. Planck, Über irreversible Strahlungsvorgänge, *Annalen d. Physik* 1, p. 69, 1900
- [54] J. Polchinski, *String Theory*, Cambridge University Press 1998
- [55] A.M. Polyakov, Conformal symmetry of critical fluctuations, *ZheTF Pis. Red.* 12, p. 538, 1970
- [56] A. Pressley, G. Segal, *Loop Groups*, Clarendon Press, Oxford 1986
- [57] K.-H. Rehren, Algebraic holography, hep-th/9905179
- [58] A. Salam, The field theory of superconductivity, *Prog.Theor.Phys.* 9, p. 550, 1953
- [59] A.N. Schellekens, On the classification of meromorphic  $c=24$  conformal field theories, *Theor. Math.Phys.* 95, p. 632, 1993
- [60] J. Scherk and J.H. Schwarz, Dual models for non-hadrons, *Nucl. Phys.* B81, p. 118, 1974
- [61] S. Schweber, *An Introduction to Relativistic Quantum Field Theory*, Harper & Row, New York 1964
- [62] J. Schwinger, ed., *Quantum Electrodynamics*, Dover, New York 1958
- [63] J. Schwinger, Euclidean quantum electrodynamics, *Phys. Rev.* 115, p. 721, 1959
- [64] J. Schwinger, Renormalization theory of quantum electrodynamics: An Individual View, in: *The Birth of Particle Physics*, Fermi Lab 1980, Proc., L.M. Brown, L. Hoddeson eds., Cambridge University Press 1983
- [65] G. Segal, Unitary representations of some Infinite dimensional groups, *Comm.math.Phys.* 80, p. 301, 1981
- [66] G. Segal, The definition of conformal field theory, in: *Differential Geometrical Methods in Theoretical Physics*, Como 1987, Proc., K. Bleuler and M. Werner eds., p. 165, NATO ASI Series 250, 1988
- [67] T.H.R. Skyrme, A non-linear theory of strong interactions, *Proc. R.Soc.* A247, p. 260, 1958; A262, p. 237, 1961
- [68] R.F. Streater, A.S. Wightman, *PCT, Spin and Statistics, and All That*, Benjamin, New York 1964
- [69] W.E. Thirring, A soluble relativistic field theory, *Ann.Phys.* 3, p. 91, 1958
- [70] M.A. Virasoro, Subsidiary conditions and ghosts in dual resonance models, *Phys. Rev.* D1, p. 2933, 1970
- [71] M.A. Virasoro, Spin and unitarity in dual resonance models, in: *Duality and Symmetry in Hadron Physics*, Proc., E. Gotsman ed., Tel Aviv 1971
- [72] J. Wess, The conformal invariance in quantum field theory, *Nuovo Cim.* 18, p. 1086, 1960

- [73] K.G. Wilson, Non-Lagrangian models of current algebra, *Phys. Rev.* 179, p. 1499, 1969
- [74] K.G. Wilson, Anomalous dimensions and the breakdown of scale invariance in perturbation theory, *Phys. Rev. D* 2, p. 1478, 1970
- [75] E. Witten, Anti de Sitter space and holography, *Adv.Theor.Math.Phys.* 2, p. 253, 1998
- [76] C.N. Yang, Einstein and his Impact on the physics of the second half of the twentieth century, in: *M. Grossmann Meeting on General Relativity, 2nd, Trieste 1979, Proc.*, R. Ruffini ed., North-Holland 1982
- [77] A.B. Zamolodchikov, Infinite additional symmetries in two-dimensional conformal quantum field theories, *Theor.Math.Phys.* 63, p. 1205, 1985

## 23. Superstring Theory – An Overview

John H. Schwarz \*

California Institute of Technology Pasadena, CA 91125, USA

### Abstract

Superstring theories and a recent extension called M theory are different facets of a unique underlying theory. They are the leading candidates for a quantum theory that unifies gravity with the other forces. As such, they are certainly not ordinary quantum field theories. However, recent duality conjectures suggest that a more complete definition of these theories can be provided by the large  $N$  limits of suitably chosen  $U(N)$  gauge theories associated to the asymptotic boundary of spacetime.

### Introduction

Superstring theory first achieved widespread acceptance during the *first superstring revolution* in 1984-85. There were three main developments at this time. The first was the discovery of an anomaly cancellation mechanism [1], which showed that supersymmetric gauge theories can be consistent in ten dimensions provided they are coupled to supergravity and the gauge group is either  $SO(32)$  or  $E_8 \times E_8$ . Any other group necessarily would give uncanceled gauge anomalies and hence inconsistency at the quantum level. The second development was the discovery of two new superstring theories—called *heterotic* string theories—with precisely these gauge groups [2]. The third development was the realization that the  $E_8 \times E_8$  heterotic string theory admits solutions in which six of the space dimensions form a Calabi-Yau space, and that this results in a 4d effective theory at low energies with many qualitatively realistic features [3]. Unfortunately, there are very many Calabi-Yau spaces and a whole range of additional choices that can be made. In any case, after the first superstring revolution subsided, we had five distinct superstring theories, each in ten dimensions. Three of them, the *type I* theory and the two heterotic theories, have  $\mathcal{N} = 1$  supersymmetry in the ten-dimensional sense. The other two theories, called *type IIA* and *type IIB*, have  $\mathcal{N} = 2$  supersymmetry [4].

The understanding of these five superstring theories was developed in the ensuing years. In each case it became clear, and was largely proved, that there are consistent perturbation expansions of on-shell scattering amplitudes. In four of the five cases (heterotic and type II) the fundamental strings are oriented and unbreakable. As a result, these theories have particularly simple perturbation expansions. Specifically, there is a unique Feynman diagram at each order of the loop expansion. The Feynman diagrams depict string world sheets, and therefore they are two-dimensional surfaces. For these four theories the unique  $L$ -loop diagram is a closed orientable genus- $L$  Riemann surface, which can be visualized as a sphere with  $L$  handles. External (incoming or outgoing) particles are represented by  $N$  points (or “punctures”) on the Riemann surface. A given diagram represents a well-defined integral of dimension  $6L + 2N - 6$ . This integral has no ultraviolet divergences, even though the spectrum contains states of arbitrarily high spin (including a massless graviton). Type I superstrings are unoriented and breakable. As a result, the perturbation expansion is more complicated for this theory, and the various world-sheet diagrams at a given order have to be combined properly to cancel divergences and anomalies.

---

\*E-mail: JHS@THEORY.CALTECH.EDU



## M Theory

In the 1970s and 1980s various supersymmetry and supergravity theories were constructed. (See [5], for example.) Ten is the largest spacetime dimension in which there exists supersymmetric Yang–Mills theories, with spins  $\leq 1$  [6], and the largest possible spacetime dimension for a supergravity theory (with spins  $\leq 2$ ), is eleven. Eleven-dimensional supergravity [7] has three kinds of fields—the graviton field (with 44 polarizations), the gravitino field (with 128 polarizations), and a three-index antisymmetric tensor gauge field  $C_{\mu\nu\rho}$  (with 84 polarizations). These massless particles are referred to collectively as the *supergraviton*. 11d supergravity, which has attracted a lot of attention over the years is nonrenormalizable, and thus it is not a consistent quantum theory. However, we now believe that it is a low-energy effective description of *M theory*, which is a consistent quantum theory [8, 9].

To explain the relation between M theory and type IIA string theory, a good approach is to identify the parameters that characterize each of them and to explain how they are related. Eleven-dimensional supergravity (and hence M theory, too) has no dimensionless parameters. The only parameter is the 11d Newton constant, or (equivalently) the 11d Planck mass  $m_p$ . When M theory is compactified on a circle (so that the spacetime geometry is  $R^{10} \times S^1$ ) another parameter is the radius  $R$  of the circle. The parameters of type IIA superstring theory are the string mass scale  $m_s$  and the dimensionless string coupling constant  $g_s$ . An important fact about all five superstring theories is that the coupling constant is not an arbitrary parameter. Rather, it is a dynamically determined as the value of a scalar field, called the *dilaton*, which is a supersymmetry partner of the graviton.

We can identify compactified M theory with type IIA superstring theory by making the following correspondences:

$$m_s^2 = 2\pi R m_p^3 \quad (1)$$

$$g_s = 2\pi R m_s. \quad (2)$$

Conventional string perturbation theory is an expansion in powers of  $g_s$  at fixed  $m_s$ . Equation (2) shows that this is equivalent to an expansion about  $R = 0$ . In particular, the strong-coupling limit of type IIA superstring theory corresponds to decompactification of the eleventh dimension, so in a sense M theory is type IIA string theory at infinite coupling. (The  $E_8 \times E_8$  heterotic string theory is also eleven-dimensional at strong coupling [10].) This explains why the eleventh dimension was not discovered in studies of string perturbation theory. These relations encode some interesting facts. For example, the fundamental IIA string actually *is* a two-dimensional membrane (called the M2-brane) of M theory with one of its dimensions wrapped around the circular spatial dimension. Denoting the string and membrane tensions (energy per unit volume) by  $T_{F1}$  and  $T_{M2}$ , one deduces that  $T_{F1} = 2\pi R T_{M2}$ . However,  $T_{F1} = 2\pi m_s^2$  and  $T_{M2} = 2\pi m_p^3$ . Combining these relations gives eq. (1). It should be emphasized that all these formulas are exact, due to the large amount of unbroken supersymmetry.

Type II superstring theories contain a variety of *p-brane* solutions that preserve half of the 32 supersymmetries [11]. These are solutions in which the energy is concentrated on a  $p$ -dimensional spatial hypersurface. (Adding the time dimension, the world volume of a  $p$ -brane has  $p+1$  dimensions.) A large class of these  $p$ -brane excitations are called *D-branes* (or *Dp-branes* when we want to specify the dimension), which have a number of special properties and are especially interesting. By definition, they are branes on which strings can end—D stands for *Dirichlet* boundary conditions. The end of a string carries a charge, and the D-brane world-volume theory contains a  $U(1)$  gauge field that carries the associated flux. When  $N$  Dp-branes are coincident, or parallel and nearly coincident, the associated  $(p+1)$ -dimensional world-volume theory is a  $U(N)$  gauge theory. The  $N^2$  gauge fields  $A_\mu^{ij}$  and their supersymmetry partners arise as the ground states of oriented strings running from the  $i$ th Dp-brane to the  $j$ th Dp-brane. The diagonal elements, belonging to the Cartan subalgebra, are massless. The field  $A_\mu^{ij}$  with  $i \neq j$  has a mass proportional to the separation of the  $i$ th and  $j$ th branes. This separation is described by the value of a corresponding scalar field in the world-volume theory.

Some D-branes have a simple M theory interpretation. In particular, the D2-brane of the type IIA theory corresponds to the M2-brane, but now in a background geometry in which one of the transverse dimensions is a circle. The first Kaluza–Klein excitation of the 11d supergraviton has mass  $1/R$ . It can be identified with the D0-brane, which accounts for eq. (2). More identifications of this type arise when we consider the magnetic dual of the M theory supermembrane. It turns out to be a five-brane, called the M5-brane. (In general, the magnetic dual of a  $p$ -brane in  $d$  dimensions is a  $(d-p-4)$ -brane.) For example, wrapping one of the M5-brane dimensions around the spatial circle gives the D4-brane.

## AdS/CFT Duality

Let me now turn to an even more recent development, which goes by the name of *AdS/CFT duality*. Here, AdS stands for *anti de Sitter space* and CFT stands for *conformal field theory*. AdS/CFT duality was proposed by Maldacena in November 1997 [13]. As is usually the case with such developments, there were important prior [14] and subsequent [15] contributions by many others.

A  $p$ -brane, or collection of  $p$ -branes, gives rise to a certain space-time geometry and gauge field configuration, which can be analyzed using the appropriate supergravity field equations. In a number of cases one finds that the geometry has an event horizon, giving a higher-dimensional analog of black holes. In some of these cases the geometry near the horizon is approximated by  $AdS_{p+2} \times S^{d-p-2}$ . This means that the AdS space has  $p+2$  dimensions and the remainder of the  $d$  dimensions form a sphere of  $d-p-2$  dimensions. There are three basic examples that have maximal supersymmetry (32 conserved supercharges). A stack of D3 branes in type IIB superstring theory has near horizon geometry  $AdS_5 \times S^5$ , a stack of M2-branes in M theory gives  $AdS_4 \times S^7$ , and a stack of M5-branes in M theory gives  $AdS_7 \times S^4$ . These solutions to type IIB and 11d supergravity were discovered in the mid 1980's, but were not pursued in the context of superstring/M theory until recently.

The basic idea of AdS/CFT duality is to identify a conformally invariant field theory (CFT) on the  $n$ -dimensional boundary with a suitable quantum gravity theory in the  $(n+1)$ -dimensional AdS *bulk*. The  $SO(2, n)$  isometries of the  $(n+1)$ -dimensional anti de Sitter space induce the group of conformal transformations on its  $n$ -dimensional Minkowski boundary. (Strictly speaking, the boundary should be compactified by adding a point at infinity.) The conformal group is therefore also  $SO(2, n)$ .

To be specific, from now on I will focus on the  $AdS_5 \times S^5$  solution of the IIB superstring theory. The IIB theory contains a four-index field  $A_{\mu\nu\rho\lambda}$  for which the D3-brane is a source. It has a field strength  $F_{\mu\nu\rho\lambda\sigma}$ , which is self-dual (in ten dimensions). In the  $AdS_5 \times S^5$  solution of the theory, the field  $F$  has a quantized flux on the sphere. Schematically,

$$\int_{S^5} F = N, \quad (3)$$

where  $N$  is a positive integer. This integer determines the radius  $R$  of the  $AdS_5$  and of the  $S^5$ , which are the same. Aside from a constant numerical factor, one finds that

$$R = (g_s N)^{1/4} \ell_s. \quad (4)$$

Thus the curvatures (which are proportional to  $R^{-2}$ ) are small compared to the string scale for  $g_s N \gg 1$  and small compared to the Planck scale for  $N \gg 1$ . The first limit suppresses stringy corrections to supergravity, whereas the latter suppresses quantum corrections to classical string theory. The conjecture is that type IIB superstring theory on  $AdS_5 \times S^5$  with  $N$  units of  $F$  flux is equivalent to  $\mathcal{N} = 4$ ,  $D = 3 + 1$   $U(N)$  Yang–Mills theory with  $g_{YM}^2 = g_s$ . For this conjecture to be plausible, it is a crucial fact the  $\mathcal{N} = 4$  super Yang–Mills theory is a CFT with vanishing beta function, a fact that was proved in the early 1980s. This duality—if true—implies an amazing fact: the 4d gauge theory, for large  $N$ , is actually a 10d string theory! Well, it is not yet “proved,” but the evidence is mounting rapidly. For example, the symmetries match: the two dual theories have the same symmetry supergroup  $SU(2, 2|4)$ .

The AdS/CFT duality conjecture has been made more precise in [15]. These papers have proposed a mapping between the bulk string theory and the boundary gauge theory. It gives a one-to-one correspondence between on-shell particles of the bulk theory and gauge-invariant operators of the boundary theory. Moreover, correlation functions of these gauge-invariant operators are related to the response of the type IIB theory to boundary conditions for the associated fields. These correspondences have been partially verified. For example, there is a perfect correspondence between particles belonging to *short* representations of the AdS supersymmetry algebra and *chiral primary operators* of the gauge theory.

The large  $N$  limit of  $SU(N)$  gauge theories for fixed  $\lambda = g_{YM}^2 N$  was studied long ago by 't Hooft [16]. He showed that only Feynman diagrams of planar topology contribute in this limit. Moreover, he conjectured that the theory should exhibit a stringy behavior in this limit. Now, this suggestion has been made precise. In principle, the complete  $\lambda$  dependence of  $\mathcal{N} = 4$  gauge theory in the 't Hooft limit should be given by *classical* type IIB superstring theory on  $AdS_5 \times S^5$ . Many people are currently studying this.

An important concept that has emerged in recent years, called the *holographic principle* [17], is incorporated by AdS/CFT duality. This concerns the number and location of degrees of freedom in a theory. In a local quantum field theory, the locality implies that the number of degrees of freedom in a spatial region is proportional to its volume. However, this cannot be correct for a quantum gravity theory, where the maximum entropy in a region is proportional to its surface area. (This bound is saturated in the case of a black hole.) So the idea of the holographic principle is that the physics in a region of space can be encoded holographically on a surface that surrounds it. This is what happens in the case of AdS/CFT duality. The physics of the AdS bulk (given by superstring theory) is not a local QFT; rather, it is projected onto the boundary theory, which is a local QFT.

## Important Unresolved Issues

One issue that needs to be settled if superstring theory is to be used for phenomenology is where supersymmetry fits into the story. It is clear that at the string scale ( $\approx 10^{18}$  GeV) the underlying theory has maximal supersymmetry (32 conserved supercharges). The question that needs to be answered is at what scales they are broken and by what mechanisms. The traditional picture (which looks the most plausible to me) is that at the compactification/GUT scale ( $\approx 10^{16}$  GeV) the symmetry is broken to  $\mathcal{N} = 1$  in  $d = 4$  (four conserved supercharges), and this persists to the TeV scale, where the final susy breaking occurs. The TeV scale is indicated by three separate arguments: the gauge hierarchy problem, supersymmetric grand unification, and the requirement that the lightest superparticle (LSP) be a cosmologically significant component of dark matter. It would be astonishing if this coincidence turned out to be a fluke. Susy GUTS are also able to account for the masses of the top and bottom quarks and the structure of electroweak symmetry breaking. Despite all these indications, we cannot be certain that this picture is correct until it is demonstrated experimentally. As I once told a newspaper reporter: discovery of supersymmetry would be more profound than life on Mars.

Another important issue is the problem of vacuum degeneracy and the stabilization of moduli. Let me explain. The underlying theory is completely unique, with no dimensionless parameters. Nevertheless, the effective potential of typical string theory quantum vacua has many flat directions, so there is a continuum, or *moduli space*, of minima. This results in parameters, called *moduli*, which characterize the vacuum values of scalar fields. These fields correspond to massless spin zero particles. Notable examples of moduli are the sizes of extra dimensions and the string coupling constant. These spin zero particles typically interact with roughly gravitational strength, which is a problem because the gravitational force is observed to be pure tensor to better than 1% accuracy. So it seems that we should seek a vacuum without moduli, which is very difficult to do. However, if a vacuum without massless scalars is ever found, it will not have any continuously adjustable parameters, and therefore it will be completely predictive (at least in principle).

Perhaps the most challenging unresolved issue of all, is the *cosmological constant*  $\Lambda$ . This is a

term in the effective action that describes the energy density of the vacuum, which is observable in a gravitational theory. Observationally, there are indications that it may be nonzero, but in any case it is extremely small compared to ordinary particle physics scales. ( $\Lambda^{1/4} \leq 10^{-11}$  GeV.) In a fundamental theory  $\Lambda$  receives contributions from many sources such as vacuum condensates and zero-point energies. Supersymmetry ensures that boson and fermion zero-point energies cancel, so the natural scale for  $\Lambda$  would seem to be the TeV susy breaking scale, which is many orders of magnitude too high. This is a fine-tuning problem that is reminiscent of the gauge hierarchy problem. Presumably string theory will provide an elegant solution. Until we know what the right mechanism is, it is hard to be confident that there is not an alternative to supersymmetry for solving the gauge hierarchy problem. I believe that when the correct solution to the problem of the cosmological constant is found, it will spark another revolution in our understanding.

## Conclusion

To conclude, there has been dramatic progress in understanding string theory in the past few years. The discovery of nonperturbative phenomena, M theory, D-branes, and AdS/CFT duality have led to many important advances including an explanation of black hole entropy and exact nonperturbative results in supersymmetric gauge theories. Further theoretical breakthroughs are still needed, however. Also, future experimental discoveries will be essential to help guide our thinking. Sooner (Tevatron or LEP) or later (LHC) exciting new phenomena are bound to show up. My bet is on Higgs and superparticles. But if I should turn out to be wrong, that would not mean that string theory is wrong.

This work supported in part by the U.S. Dept. of Energy under Grant No. DE-FG03-92-ER40701.

## References

- [1] Green, M.B., and Schwarz, J.H., *Phys. Lett.* **149B** (1984) 117.
- [2] Gross, D.J., Harvey, J.A., Martinec, E., and Rohm, R., *Phys. Rev. Lett.* **54**, (1985) 502.
- [3] Candelas, P., Horowitz, G.T., Strominger, A., and Witten, E., *Nucl. Phys.* **B258** (1985) 46.
- [4] Green, M.B., and Schwarz, J.H., *Phys. Lett.* **109B** (1982) 444.
- [5] Salam, A., and Sezgin, E., eds., *Supergravities in Diverse Dimensions*, reprints in 2 vols., World Scientific (1989).
- [6] Brink, L., Schwarz, J.H., and Scherk, J., *Nucl. Phys.* **B121** (1977) 77; Gliozzi, F., Scherk, J., and Olive, D., *Nucl. Phys.* **B122** (1977) 253.
- [7] Cremmer, E., Julia, B., and Scherk, J., *Phys. Lett.* **76B** (1978) 409.
- [8] Townsend, P.K., *Phys. Lett.* **B350** (1995) 184, hep-th/9501068.
- [9] Witten, E., *Nucl. Phys.* **B443** (1995) 85, hep-th/9503124.
- [10] Hořava, P., and Witten, E., *Nucl. Phys.* **B460** (1996) 506, hep-th/9510209.
- [11] Horowitz, G.T., and Strominger, A., *Nucl. Phys.* **B360** (1991) 197.
- [12] Polchinski, J., *Phys. Rev. Lett.* **75** (1995) 4724, hep-th/9510017.
- [13] J.M. Maldacena, *Adv. Theor. Math. Phys.* **2** (1998) 231, hep-th/9711200.
- [14] I.R. Klebanov, *Nucl. Phys.* **B496** (1997) 231, hep-th/9702076; S.S. Gubser, I.R. Klebanov, and A.A. Tseytlin, *Nucl. Phys.* **B499** (1997) 217, hep-th/9703040; S.S. Gubser and I.R. Klebanov, *Phys. Lett.* **B413** (1997) 41, hep-th/9708005; A.M. Polyakov, hep-th/9711002.
- [15] S.S. Gubser, I.R. Klebanov, and A.M. Polyakov, *Phys. Lett.* **B428** (1998) 105, hep-th/9802109; E. Witten, *Adv. Theor. Math. Phys.* **2** (1998) 253, hep-th/9802150.
- [16] G. 't Hooft, *Nucl. Phys.* **B72** (1974) 461.
- [17] G. 't Hooft, gr-qc/9310026; L. Susskind, *J. Math. Phys.* **36** (1995) 6377, hep-th/9409089.

# 24. Recent Developments In String Theory

Jnanadeva Maharana \*

Institute of Physics Bhubaneswar - 751005 India

## Abstract

The purpose of this short review is to present progresses in string theory in the recent past. There have been very important developments in our understanding of string dynamics, especially the nonperturbative aspects. In this context, dualities play a cardinal role. The string theory provides a deeper understanding of the physics of special class of black holes from a microscopic point of view and has resolved several important questions. It is also recognized that M-theory provides a unified description of the five perturbatively consistent string theories. The article covers some of these aspects and highlights important progress made in string theory.

## Contents

1. Introduction	612
2. Perturbative Aspects of String Theory	615
3. Duality Symmetries in String Theory	624
4. M-theory and Unified String Dynamics	637
5. Black holes and String Theory	642
6. M-theory and the Matrix model	647
7. Anti-de Sitter Space and Boundary Field Theory Correspondence	652
8. Summary and Conclusions	658
9. References	659

---

\*Email: maharana@iopb.res.in

# 1 Introduction

All along the progress in natural philosophy, curious minds have asked deep questions pertaining to the fundamental constituents of matter and creation and evolution of the cosmos. In the modern era, physicists have endeavored to comprehend natural phenomena in terms of a simple set of principles. Therefore, the search has continued to discover the elementary constituents of matter and identify the fundamental forces responsible for the natural phenomena. It is accepted that there are four fundamental forces : gravitation, the weak interaction, electromagnetism and the strong interaction. The unification of fundamental interactions has remained as one of the most outstanding challenge for generations of physicists. In the latter half of this century, some progress has taken place in this direction through the electroweak unification scheme. The electroweak theory together with quantum chromodynamics (QCD), referred to as the standard model, have been tested to a great degree of accuracy. Thus, the standard model provides a very good description of the 'low energy physics', comprising of the spectrum of elementary particles and their dynamics. The next step in fulfilling the dream of unification of forces were the schemes of grand unifications (GUT) which attempted to incorporate the three fundamental interactions, leaving aside gravitational interaction. The QED has been tested to a great degree of accuracy and two most important characteristics of that theory are the invariance under local gauge transformations and renormalizability. The electroweak theory and QCD respect the principle of gauge invariance and are renormalizable. Moreover, it is well known that the Einstein's theory of general relativity respects a local symmetry: invariance under general coordinate transformations. However, the theory is not renormalizable since the Newton's constant carries dimension of  $(mass)^{-2}$ , unlike the gauge coupling constants of the standard model which are dimensionless.

Although the standard model has successfully passed many stringent experimental tests, it is recognised that one must seek for a more fundamental theory. The standard model has many arbitrary parameters: the gauge coupling constants, the coupling constants of the scalars, Yukawa couplings of the Higgs bosons to fermions which are eventually responsible for generating fermion masses, just to mention a few. Furthermore, when one extrapolates the gauge coupling constants utilizing the renormalization group equations towards higher energy scale, there are evidences that the three coupling constants tend to converge to a point and it is natural to conclude that beyond that scale there might be a unified description of the standard model. Therefore, these observations lend support to the proposal of GUTS put forward in early seventies. As is well known, the existence of electroweak scale in the TeV region and another unification scale in the neighbourhood of  $10^{16}$  to  $10^{17}$  GeV leads to issues related to fine tuning of parameters, known as gauge hierarchy problem. The gauge hierarchy problem can be resolved in an elegant manner if one envisages supersymmetric version of the standard model (moreover, the convergence of gauge coupling constants in the unifying scale is more favourable in supersymmetric theories; see Mohapatra's article in this volume for details). The supersymmetric theories were constructed so that bosons and fermions can belong to a supermultiplet. The supersymmetry appeared in 2-dimensions in the construction of string theories. While attempts were being made to construct various types of grand unified theories, there were developments in incorporating gravity into supersymmetric theories which resulted in discovery of supergravity theories. However, it was not possible to construct renormalizable field theories which could unify the four fundamental forces. It was being perceived by many physicists, in the beginning of eighties, that new radical ideas were required to unify the fundamental interactions.

It is now accepted that string theory holds the promise of unifying all the fundamental interactions. The progress of the string theory in diverse directions, during the last fifteen years have been truly spectacular. The theory has not only brought us nearer to the dream of unification, but also has influenced our understanding of various aspects of quantum gravity and has revealed many beautiful features relevant to the nonperturbative aspects of field theories.

The string theory was invented to describe the dynamics of strongly interacting particles. The vast amount of experimental data amassed from high energy accelerators, during fifties and sixties led to discovery of large number of hadronic resonances. One of the interesting characteristics of those resonances was that when one plots squared of mass vs spins of these particles, families of the res-

onances tend to lie on a straight line, known as the Chew-Fraustchi plot. It was also evident from the high energy of scattering cross sections of hadrons that they follow a power law behaviour i.e. the crossed channel Regge poles controlled cross sections at high energies. The duality relation in strong interactions, that is sum over direct channel resonances (from low energy data) reproduces the Regge amplitude, was an important discovery for construction of dual models. Veneziano [1] took crucial step when he proposed a four point amplitude which satisfied requirements of duality and crossing symmetry.

$$T(s, t) = B(s, t) + B(t, u) + B(u, s) \quad (1.1)$$

where

$$B(s, t) = \frac{\Gamma(-\alpha(s))\Gamma(-\alpha(t))}{\Gamma(-\alpha(s) - \alpha(t))} \quad (1.2)$$

and  $\alpha(s) = \alpha_0 + \alpha's$  is the parametrization of the linear Regge trajectory. Here s, t and u refer to the Mandelstam variables; when we are in the center of mass frame, s is the squared of c.m. energy, t and u are related to the c.m. scattering angle. The B-function has the integral representation

$$B(s, t) = \int_0^1 dw w^{-1-\alpha(s)} (1-w)^{-1-\alpha(t)} \quad (1.3)$$

Subsequently, generalized N-point amplitudes satisfying requirements of duality and crossing symmetry were proposed by several authors [2] and one such amplitude is

$$F(p_1, \dots, p_N) = |w_I - w_{II}| |w_{II} - w_{III}| |w_{III} - w_I| \int dw_1 \dots dw_N \prod_{i < j} |w_i - w_j|^{2\alpha' p_i \cdot p_j} \quad (1.4)$$

$w_i$  are ordered cyclically,  $w_I, w_{II}$  and  $w_{III}$  are any three of the variables of the set  $\{w_i\}$ , but are held fixed. As in the case of 4-point Veneziano amplitude, the full N-point amplitude is sum of all cyclically inequivalent permutations. It was realized that it is possible to represent the N-point function in a path integral form [3]

$$F_N \sim \int \Pi_{\mu, \sigma} dX^\mu(\sigma) \int dw_1 \dots dw_N \exp\left(-\frac{T}{2} \int_{\sigma^2 > 0} d^2\sigma \partial_a X^\mu \partial^a X^\nu \eta_{\mu\nu}\right) \Pi^N e^{ip_I \cdot X(w_I)} \quad (1.5)$$

where  $\partial_a X^\mu = \frac{\partial X^\mu}{\partial \sigma^a}$ ,  $\sigma^1$  and  $\sigma^2$  are coordinates of a point in the upper plane,  $X^\mu(\sigma)$  are integrated over all functions of  $\sigma$ . The boundary condition on  $X^\mu$  is  $\partial_2 X^\mu = 0$  for  $\sigma^2 = 0$ . The constant  $T = \frac{1}{2\pi\alpha'}$  was later on identified as the tension of the string. Note the presence of  $X^\mu(w_I)$ ; it is the value of  $X^\mu(\sigma^1, \sigma^2)$  on the line  $\sigma^1 = w, \sigma^2 = 0$ . The connection between dual amplitudes and dynamics of a relativistic string was recognised by several authors independently [4]. Now, of course we know that this amplitude is obtained from an open string theory and the action is that of a string, there are vertex operator in the path integral formula and the open string boundary conditions are to be specified. Virasoro had constructed another 4-point amplitude [5] fulfilling the requirement of duality and crossing symmetry in sequel to Veneziano's paper and generalization of that amplitude for N-particle scattering was derived with a path integral representation [6]. It was realized that the Virasoro-Shapiro amplitude could be obtained from a closed string theory. Finally, Nambu proposed the action for the string so that one could start studying the dynamics of the string and proceed to examine the consequences of its quantization.

The string theory as a theory of strong interaction dynamics was not free from shortcomings. While attempts were going on to rectify the pitfalls of the theory and to construct new string theories as models of strong interactions; QCD was proposed as the fundamental theory of strongly interacting particle. The theory described interactions of the fundamental constituents, quarks, of the hadrons with gluon as the carrier of the force. Furthermore, the experimental data confirmed predictions of QCD steadily and consequently; string theory as a theory of strong interaction was no longer in the center stage.

In 1974, Joel Scherk and John Schwarz [7] made a bold proposition that string theory should be envisaged as a theory of gravity since the massless spin two particle appears naturally in the closed



string spectrum and this theory might be a vehicle to achieve the goal of unification of the forces of Nature. If string theory were to incorporate the gravitational interaction, then the string tension should be order of the Planck scale in contrast to the the tension of the original string which was of the order of one GeV, the scale of hadronic interaction determined from the slope of the Regge trajectories. It was realised that one has to go up nineteen orders of magnitude in the energy scale if the Scherk-Schwarz proposal was to be realized. At that time, this radical idea did not receive wide spread acceptance amongst theoretical high energy physicists. The crucial work of Green and Schwarz [8] in the summer of 1984 led to conclusion that 10-dimensional super Yang-Mills theories coupled to supergravity can be consistently constructed and are free from all anomalies [9] only for the gauge groups  $SO(32)$  and  $E_8 \times E_8$ . The results of Green and Schwarz had profound impact on the field of high energy physics. It was recognised that string theory could fulfill the cherished dream of unifying fundamental forces. The construction of the heterotic string theory [10] was a very important break through towards realization of this goal since it had the desired gauge groups i.e.  $SO(32)$  or  $E_8 \times E_8$ , depending on the construction one adopted. The ten dimensional theory had chiral fermions,  $N = 1$  supergravity coupled to supersymmetric Yang-Mills with appropriate gauge groups. Moreover, when the  $E_8 \times E_8$  heterotic string theory was compactified to four dimensions on a Calabi-Yau manifold, the resulting theory was shown to possess several desirable features that one expected from some of the grand unified theories. Furthermore, it was possible to demonstrate that the standard model gauge groups  $SU(3) \times SU(2) \times U(1)$  were contained in such four dimensional theories. Indeed, optimistically, one could feel that a unified theory was in sight and string theory was popularly named as the ‘Theory of Everything’.

Let us recapitulate some of the essential features of string theory. The string is a one dimensional object which executes motion in spacetime. There are, grossly speaking, two types of strings: open and closed strings. As the name suggests, the ends of open strings are free (there are special types whose ends might get stuck to some surfaces and they play very important roles too) and it is required to satisfy suitable boundary conditions for the end points. The closed string, by definition, has its both ends glued together, forming a loop. It is well known that when a point particle evolves in spacetime, it traces out a trajectory describing its history. In case of an open string, it sweeps a two dimensional surface and similarly the closed string sweeps a surface which is that of a cylinder. The natural question is why we do not observe these strings in high energy collisions. The answer to this question lies in the fact that the strings are much smaller in size than the present accelerators can probe. If we could have accelerators which have energies of the order of  $10^{19}$  GeV, then it will be possible to observe the dynamics of the strings directly and test the predictions of string theory at the Planckian energies. In contrast, the present day accelerators have energies of the order of TeV - almost 16 orders of magnitudes below the string scale.

The string has tension and it vibrates in an infinite number of modes. We identify each mode of the string with a particle. Of course, the string will have the lowest mode and we identify that with a particular particle. The next mode will correspond to an excited state and it is separated in energy from the lowest mode in suitable unit of string tension - separation between two neighbouring levels is order of  $10^{19}$  GeV (recall that for the hadronic models they excitations were on Regge trajectories and there the tension was order of GeV). The string theories of interests to us contain massless particles in their lowest mode. For example, in 10-dimensional heterotic string theory, we have graviton, antisymmetric tensor and dilaton together with the super Yang-Mills multiplets corresponding to the gauge groups  $SO(32)$  or  $E_8 \times E_8$  in its massless sector. Therefore, in the low energy limit, the string theory effectively reduces to a point particle field theory (this is when we want to describe physics at the present day accelerator energy scales). In other words, the zero slope limits of string theories correspond to known field theories - superstring theories go over to supergravity theories in this limit.

Now we give an outline of the rest of the article. Since it is to appear in a volume on ‘Field Theory’, we shall avoid involved technical details. The field has progressed in diverse directions and our strategy will be to adopt a course to high light important developments. We shall attempt to present different aspects of string theory in a pedagogical manner. In order to get across some issues, known examples from field theory will be presented. There has been intense activities in this field since 1984, when it was recognised that string theory is the most promising candidate for

unification of forces of Nature. It is not possible to cover all the important literatures in a vibrant field like this within the frame work of this article. I apologize in advance to all the authors whose works have not been cited. There are two books which cover all the important aspects of string theory in detail besides several monographs and reprint collection volumes. The first one [11], in two volumes provides foundation for string theory and includes the developments up to 1986. The second one [12] has laid the emphasis on the progress made after the second superstring revolution. I have listed some of the review articles written in the first phase of the developments of string theory [13, 14, 15, 16, 17, 18]. There are a large number of review articles written in recent time [19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]. The next section deals with a brief review of the perturbative aspects of string theory to familiarize the reader with well known results. First, the string worldsheet action is introduced and the symmetries of the action are listed. A very quick exposition is given to the solutions of the equations of motion and mode expansions for the string coordinates and essentials of the Virasoro algebra are recalled. The evolution of the string in its massless background in the first quantized approach is discussed and the consequences of conformal invariance are noted. Section III deals with the symmetries of string theory. The theory is endowed with a rich symmetry structure in the target space besides the worldsheet symmetries. We introduce the duality symmetries since they play a very important role in our understanding of the string dynamics in various spacetime dimensions and they unravel the intimate connections between different string theories. The subsequent Section, IV, is devoted to to discuss the recent efforts to unify string theories. Besides duality symmetry, spatially extended objects, generically called p-branes, which appear as solutions to the effective action, are crucial to our understanding of string dynamics and to test some of the duality conjectures. We introduce some of the salient features of these objects and provide simple examples of the solutions. The *raison de etre* for M-theory is presented. We give an example how compactification of M-theory provides connections with the string theories and their various brane content. The fifth Section deals with issues related to black holes that appear in string theory. Since string theory describes gravity, it is expected that the theory will be able to provide insights into deep questions in quantum gravity. Indeed, some of the issues in the physics of the black holes have been resolved by string theory. It is known for more than two decades that a black hole is characterized by entropy and the Hawking temperature from the thermodynamic analogies. Moreover, the seminal work of Hawking demonstrated that the black holes radiate when quantum effects are taken into account. Recently, the black hole entropy has been computed as a microscopic derivation in the frame work of string theory. Furthermore, the absorption cross sections for incident particles and the distribution of Hawking radiation emitted from special class of stringy black holes have been evaluated from a microscopic theory. Section VI contains a brief account of the M(atr)ix model. The M(atr)ix model proposal to describe M-theory has drawn considerable attention. Some of the calculations in this model give surprising agreements with results of supergravity theories. Moreover, when one considers compactification of the model on torii the resulting theory can be related to supersymmetric Yang-Mills theories through duality. We discuss some of the features of the Maldacena conjectures in Section VII. According to the conjecture, in a concrete form, if one considers  $N$  coincident 3-branes of type IIB theory on  $AdS_5 \times S^5$  then the correlation functions of supergravity on the  $AdS_5$  get related to correlation functions of the  $N = 4$  super Yang-Mills theory living on the boundary of the  $AdS_5$ . This is a rapidly developing area and we shall be contented with some of the simple examples. In the last section we present an overview of the field. We make a few remarks to convey the reader how the work in string theory has influenced research in other branches of physics.

## 2 Perturbative Aspects of String Theory

We have outlined the historical backgrounds and the developments of string theory in its early phase in the previous section. In this section, we shall present some of the essential features of string theory such as its quantization, the perturbative spectrum of theory and the supersymmetric version of string theory.

Nambu had proposed the action for a string in analogy with the action for a relativistic point

particle: the action for point particle in an integral over an line element; the string action is expected to be an integral over a surface.

The Nambu-Goto action [34] was introduced almost three decades ago and has the form

$$S_{NG} = -T \int d^2\sigma \sqrt{(\dot{X}.X')^2 - (\dot{X})^2(X')^2} \quad (2.1)$$

where  $\sigma$  and  $\tau$  are the coordinates on the surface swept out by the the string, called 'worldsheet';  $\dot{X}^\mu = \frac{\partial X^\mu}{\partial \tau}$  and  $X'^\mu = \frac{\partial X^\mu}{\partial \sigma}$  and we shall follow this definition all along unless specified otherwise. The equations of motion can be derived after specifying boundary conditions for the types of string one is dealing with. One important point to be noted is that the theory described by the above action satisfies two constraints

$$\Pi.X' = 0, \quad \Pi^2 + TX'^2 = 0 \quad (2.2)$$

where  $\Pi_\mu = \frac{\delta L}{\delta \dot{X}^\mu}$  is the canonical momentum of  $X^\mu$  obtained from this action. We reserve the notation  $P_\mu$  for the canonical momentum of the coordinate derived from the Polyakov action. We shall elaborate significance of these constraints later.

However, this form of action was not very convenient to deal with the quantization of string and an alternative form of action was proposed by Polyakov [35]

$$S = -\frac{T}{2} \int d^2\sigma \sqrt{-\gamma} \gamma^{ab} \partial_a X^\mu \partial_b X^\nu \eta_{\mu\nu} \quad (2.3)$$

$\gamma_{ab}$  is the worldsheet metric,  $\gamma^{ab}$  is its inverse,  $\gamma$  is determinant of worldsheet metric and  $\eta_{\mu\nu}$  is the flat space metric of the target space. The variation of the action with respect to  $\gamma^{ab}$  results in the worldsheet stress energy momentum tensor

$$T_{ab} = \partial_a X \cdot \partial_b X - \frac{1}{2} \gamma_{ab} \gamma^{cd} \partial_c X \cdot \partial_d X \quad (2.4)$$

$T_{ab} = 0$ , since there is no kinetic term i.e. as the analogue of Einstein-Hilbert piece,  $\int d^2\sigma R^{(2)}$  is a topological term. We can solve for  $\gamma_{ab}$  from the above equation

$$\gamma_{ab} = \partial_a X^\mu \partial_b X^\nu \eta_{\mu\nu} \quad (2.5)$$

If we insert the above expression for worldsheet metric into the Polyakov action, then we recover Nambu's action.

The action (2.3) has following symmetry properties.

(a) Two dimensional reparemetrization invariance

$$\delta\gamma_{ab} = \xi^c \partial_c \gamma_{ab} + \partial_a \xi^c \gamma_{bc} + \partial_b \xi^c \gamma_{ac} \quad (2.6)$$

and hence  $\delta\sqrt{-\gamma} = \partial_a(\xi^a \sqrt{-\gamma})$ . The string coordinate transforms as

$$\delta X^\mu = \xi^a \partial_a X^\mu \quad (2.7)$$

Weyl invariance

$$\delta\gamma_{ab} = 2\Omega\gamma_{ab}, \quad \delta X^\mu = 0 \quad (2.8)$$

Poincare invariance (in target space)

$$\delta X^\mu = \omega_\nu^\mu X^\nu + a^\mu, \quad \delta\gamma_{ab} = 0 \quad (2.9)$$

where  $\omega_\nu^\mu$  are antisymmetric parameters associated with the Lorentz transformation and  $a^\mu$  are the parameters of translation.

Note that the Weyl invariance implies tracelessness of the two dimensional energy momentum tensor for the classical theory. The quantum invariance of this symmetry has far reaching consequences in string theory.

If we make the orthonormal gauge choice for the worldsheet metric  $\gamma_{ab} = e^{2\Omega(\sigma,\tau)}\eta_{ab}$  with  $\eta_{ab} = \text{diag}(-1, +1)$  the, form of Polyakov action simplifies since  $\sqrt{-\gamma}\gamma^{ab} = \eta^{ab}$  in this gauge. The condition of the vanishing of  $T_{ab}$  reduces to two constraints

$$(\dot{X} \pm X')^2 = 0 \quad (2.10)$$

These are the Virasoro constraints. They take the following form in the Hamiltonian formalism

$$P_\mu X'^\mu = 0, \quad H = \frac{1}{2}(P^2 + T X'^2) = 0 \quad (2.11)$$

where  $P_\mu$  is momentum conjugate to  $X^\mu$  derived from Polyakov action. It is easy to see that the first constraint generates  $\sigma$  translation on the worldsheet, whereas latter being the canonical Hamiltonian generates  $\tau$  translation.

The equation of motion for the string coordinates, in the light-cone variables  $\xi_+ = \tau + \sigma$  and  $\xi_- = \tau - \sigma$ , are given by

$$\partial_+ \partial_- X^\mu = 0 \quad (2.12)$$

We note that the equation of motion is derived with following boundary conditions: (i)  $X^\mu(\tau, \sigma + 2\pi) = X^\mu(\tau, \sigma)$  for the closed strings, and (ii)  $X'^\mu = 0$  for  $\sigma = 0$  and  $\sigma = 2\pi$  in the case of open strings, when we apply the variational method.

Let us illustrate the mode expansion for the closed string starting from the equation of motion with periodic boundary condition in  $\sigma$ . We first note that the string coordinate can be decomposed as a sum of left-moving and right-moving coordinates.

$$X^\mu(\tau, \sigma) = X_L^\mu(\tau + \sigma) + X_R^\mu(\tau - \sigma) \quad (2.13)$$

Then the two can be expanded as follows:

$$X_L^\mu(\tau + \sigma) = \frac{x^\mu}{2} + \frac{p^\mu}{4\pi T}(\tau + \sigma) + \frac{i}{\sqrt{4\pi T}} \sum \frac{\bar{\alpha}_m^\mu}{m} e^{-im(\tau + \sigma)} \quad (2.14)$$

$$X_R^\mu(\tau - \sigma) = \frac{x^\mu}{2} + \frac{p^\mu}{4\pi T}(\tau - \sigma) + \frac{i}{\sqrt{4\pi T}} \sum \frac{\alpha_m^\mu}{m} e^{-im(\tau - \sigma)} \quad (2.15)$$

The sum is over all integer values of  $m$  ( $m = 0$  is excluded) in the above equations.  $\alpha_m^\mu$  and  $\bar{\alpha}_m^\mu$  are the Fourier modes. Since  $X_{L,R}^\mu$  are real, so are  $x^\mu$  and  $p^\mu$ ; the Fourier modes satisfy

$$(\alpha_m^\mu)^* = \alpha_{-m}^\mu, \quad (\bar{\alpha}_m^\mu)^* = \bar{\alpha}_{-m}^\mu \quad (2.16)$$

from the reality condition. For the closed string case, the classical Hamiltonian is given by

$$H = \frac{1}{2} \left( \sum \alpha_m \cdot \alpha_{-m} + \sum \bar{\alpha}_m \cdot \bar{\alpha}_{-m} \right) \quad (2.17)$$

in terms of the Fourier modes. The constraint,  $T_{ab} = 0$ , obtained from the Polyakov action, takes the form  $T_{--} = \frac{1}{2}(\partial_- X)^2 = 0$  and  $T_{++} = \frac{1}{2}(\partial_+ X)^2 = 0$ , in terms of the light-cone coordinates, after one has gone over to the ON gauge. It is more convenient to express these constraints in terms of the Fourier modes introduced above and define the Virasoro generators

$$L_m = \frac{1}{2} \sum \alpha_{m-n} \cdot \alpha_n, \quad \text{and} \quad \bar{L}_m = \frac{1}{2} \sum \bar{\alpha}_{m-n} \cdot \bar{\alpha}_n \quad (2.18)$$

And

$$H = L_0 + \bar{L}_0 \quad (2.19)$$

We can obtain classical Poisson bracket relations amongst  $L_m$ , similarly for the set  $\bar{L}_m$ , starting from the canonical Poisson bracket between  $X^\mu$  and  $P_\mu$ .

$$[L_m, L_n]_{PB} = -i(m - n)L_{m+n} \quad (2.20)$$

$$[L_m, \bar{L}_n]_{PB} = 0 \quad (2.21)$$

the PB between  $\{\bar{L}_m\}$  is same as for  $\{L_m\}$ . These are classical Virasoro algebra.

We have noted earlier that the string theory is endowed with local symmetries in the worldsheet and the action is that of D-scalar fields in  $1 + 1$  dimensions, since  $\mu = 0, 1 \dots D - 1$  takes D values. When we proceed to quantize this theory, we encounter problems similar to the one faced in quantization of gauge theory. In other words we have to fix the gauge here too. One can choose to work in a noncovariant gauge which has the advantage of dealing with physical degrees of freedoms directly, but at the price of losing manifest Lorentz covariance. On the other hand one can adopt covariant quantization prescription with all its elegance and power. The light-cone quantization, although noncovariant, is very useful and gives us a physical picture. As the first step, the classical constraints are solved and one is left with less number of variables. Recall that there were some remnant symmetries after choosing conformal (ON) gauge:  $\xi'_+ = \lambda_1(\xi_+)$  and  $\xi'_- = \lambda_2(\xi_-)$ . One can utilise this property to write

$$X^+ = x^+ + \alpha' p^+ \tau \quad (2.22)$$

Defining the light-cone string coordinated  $X^\pm = X^0 \pm X^1$  one can impose the classical Virasoro constraints  $(\dot{X} \pm X')^2 = 0$ . Thus  $X^-$  is determined in terms of the rest of the (transverse) coordinates,  $X^i$ ; and in this process both  $X^+$  and  $X^-$  are totally eliminated and we are left with  $\{X^i\}$ . Then the oscillators of these coordinates will create the states which could be identified with particles with physical degrees of freedom only. So it gives us a physical picture of the states. However, as mentioned earlier and as is the case with noncovariant gauge fixing in QED or Yang-Mills theories, the Lorentz invariance must be checked explicitly. For the case of string theory, one is required to construct the generators of Lorentz transformations and ensure that the generators satisfy the algebra. It is well known that this requirement is not fulfilled unless the string propagates in 26-dimensional spacetime. On the other hand, if one adopts the covariant BRST procedure, it is necessary to add the ghost term to the action and construct the corresponding Virasoro generators for the ghosts. Thus the full Virasoro generator is a sum of the oscillators coming from string coordinates and those from the ghosts. When we compute the quantum Virasoro algebra, there is an anomaly of 26 from the ghost sector which gets precisely canceled if the spacetime dimension is 26 since each bosonic degrees of freedom contributes a factor of one to the anomaly with a sign just opposite to that coming from ghosts.

There are infinite tower of states in string theory. It is useful to arrange them according to their oscillator levels. Notice that the worldsheet degrees of freedom of the string are envisaged as a collection of infinite number of harmonic oscillators. If we consider creation operator of one of these oscillators, we could define a level such that the number of units of worldsheet momenta created by this operator while acting on the vacuum. If we have a state, then the total oscillator level of that state is the sum of the levels of all the oscillators acting on the Fock vacuum to create this state. For the free string, the coordinates can be decomposed into left moving and right moving sectors. Therefore, one can define left and right moving oscillator levels (same decomposition is valid when we add fermionic degrees of freedom). Thus one can write  $L_0 = \frac{1}{2}(E + P)$  and  $\bar{L}_0 = \frac{1}{2}(E - P)$ , where E and P are worldsheet energy and momentum respectively. Therefore,  $L_0$  and  $\bar{L}_0$  get contributions from the oscillators and from the Fock vacuum. We may remark in passing that the momenta of the spacetime D-dimensional theory ( $25 + 1$  for bosonic string and  $9 + 1$  for superstring) are the ones conjugate to the zero modes of the bosonic/and/or fermionic worldsheet theory. Therefore, the ground state of the closed bosonic string is a tachyon satisfying the relation  $\alpha' m^2 = -4$ , with  $\alpha' = \frac{1}{2\pi T}$ . The first excited (massless) states of closed string are:

- Spin 2 state,  $G_{\mu\nu}$ , identified as graviton.
- An antisymmetric tensor field,  $B_{\mu\nu}$ .
- A scalar,  $\phi$ , called dilaton.

They belong to the irreducible representation of the  $SO(24)$  group. These states are created by action of a single creation operator from the left moving sector and another creation operator from the right moving sector. Therefore, they will have two target space Lorentz indices and one can decompose them according to irreducible representations of the corresponding rotation group.

Introduction worldsheet fermions has important consequences. In fact, if one demands worldsheet superconformal symmetry generalising from the bosonic string coordinates to include fermionic degrees of freedom, then resulting theory is the superstring. First we need to construct two dimensional supergravity action. One needs to add to the action (2.3) the action

$$-\frac{T}{2} \int d^2\sigma e \{ i\psi^\mu \gamma^0 \gamma^a \partial_a \psi_\mu - i\bar{\lambda}_a \gamma^b \gamma^a \psi^\mu \partial_b \psi_\mu - \frac{1}{4} \psi^\mu \gamma^0 \psi_\mu \bar{\lambda}_a \gamma^b \gamma^a \lambda_b \} \quad (2.23)$$

The notations are as follows [37]:  $\psi^\mu$  are worldsheet two component Majorana fermions,  $e_a^i$  are the zweibeins associated with the worldsheet metric,  $e$  is its determinant.  $\lambda^a$  is the gravitino on worldsheet satisfying  $\lambda_a^* = \lambda_a$ . The gamma matrices in the worldsheet have following representations:  $\gamma^0 = \sigma_2, \gamma^1 = i\sigma_1$  and  $\gamma_5 = \gamma^0 \gamma^1 = \sigma_3$ ,  $\sigma_i$  being the three Pauli matrices. We shall go over to the superorthonormal gauge, where the worldsheet metric is flat metric times a conformal factor (mentioned already) and gravitino is chosen to be  $\lambda_a = \gamma_a \zeta$  where  $\zeta$  is a constant Majorana spinor. Then the action (2.23) takes a simple form and is expressed in terms of the Weyl Majorana fermions (it is a free fermion theory now)

$$-\frac{iT}{2} \int d^2\sigma [\psi_+^\mu (\partial_\tau - \partial_\sigma) \psi_{+\mu} + \psi_-^\mu (\partial_\tau + \partial_\sigma) \psi_{-\mu}] \quad (2.24)$$

with the definition of the chiral fermions:  $\psi_+ = \frac{1}{2}(1 - \gamma_5)\psi$  and  $\psi_- = \frac{1}{2}(1 + \gamma_5)\psi$ , the space-time index is suppressed. Now it is evident that the fermion equations of motion will separated according to the chiralities, as is expected for massless fermions. The worldsheet supersymmetry transformations are

$$\delta X^\mu = \bar{\epsilon} \psi^\mu \quad (2.25)$$

$$\delta \psi^\mu = -i\gamma^a \partial_a X^\mu \epsilon \quad (2.26)$$

For the two component Majorana fermions;  $\epsilon$  is the fermionic parameter associated with the supersymmetry transformation. The supercharge is the time component of supercurrent integrated over  $\sigma$  variable. The current is

$$J^a = \gamma^b \partial_b X^\mu \gamma^a \psi_\mu \quad (2.27)$$

Next, one defines the super Virasoro generators and compute the quantum algebra and derive the condition for absence of anomaly. In case of the superstring the critical dimension is ten in contrast to bosonic string where it was 26.

Now we shall consider a few points before discussing how spacetime supersymmetry multiplets appear in the spectrum of the superstring. We had mentioned that the bosonic string has a tachyon in its lowest level which will render the theory unstable. Although, worldsheet supersymmetric theory moves in ten dimensional spacetime, the super Virasoro algebra does not impose sufficient constraints to remove the undesirable tachyon from the spectrum in general.

Notice from the fermionic equations of motion (we suppress the bosonic part momentarily to focus attentions on fermions only) that there is some freedom in the choice of the boundary condition as  $\sigma$  goes over a period of  $2\pi$ . This is due to the fact that the action remains invariant under  $\psi \rightarrow -\psi$  for fermions of either chirality. The boundary conditions are:

$$\psi(\sigma + 2\pi) = -\psi(\sigma) \quad (2.28)$$

known as Neveu-Schwarz boundary condition is antiperiodic[38]. The periodic boundary condition

$$\psi(\sigma + 2\pi) = \psi(\sigma) \quad (2.29)$$

is the Ramond condition [39]; the indices are suppressed for notational convenience. The mode expansion for, say the holomorphic field, is

$$\psi_+^\mu(\tau + \sigma) = \sum_n \psi_n^\mu e^{-n(\tau + \sigma)} \quad (2.30)$$

It is easy to see that for Ramond boundary condition,  $n$  must be integers. When we impose NS (Neveu-Schwarz) boundary condition and expand the fermions in Fourier modes, then  $n$  will take

half integer values. We note that the NS fermions have no zero modes, whereas the Ramond fermions have zero modes in the Fourier expansions.

Let us extend the arguments, we used for the bosonic string, for the superstring and examine their spectrum. The aim is to get rid of the tachyon and to construct states using bosonic and fermionic operators such that these states transform like fermions and the resulting theory be endowed with spacetime supersymmetry. We shall consider the light-cone gauge so that physical degrees of freedom become transparent. In addition to the condition  $X^+ = x^+ + p^+ \tau$ , one imposes constraint

$$\psi^+ = \bar{\psi}^+ = 0 \quad (2.31)$$

for the NS fermions, when we have Ramond fermions, they can be set to zero except for the zero modes. Now we look at the superconformal constraints and solve for  $X^-, \psi^-, \bar{\psi}^-$  in terms of the rest of the coordinates. Thus we can use the (physical) transverse oscillators of both  $X$  and  $\psi$  to construct the physical states and keep in mind the presence of appropriate zero modes. It follows from straight forward calculation that the ground state in the NS sector is tachyon. The next level obtained by operating  $\psi^i$  contains massless states. Thus we need to remove the tachyon as well as some of the unwanted states, at the same time, keeping the massless spectrum in tact. Notice that worldsheet fermions are anticommuting objects, although they create bosonic states while operating on a state of the theory. This feature is not very desirable as will be evident from the following example. Let us consider a specific bosonic state of a superstring and then operate on it a worldsheet spinor,  $\psi_+^i$ , obeying NS boundary condition. The resulting state will still be an integer spin object even if we have operated by anticommuting operator; this is rather unusual. We can think of a situation when odd number of NS operators act on a bosonic state and obviously same situation will continue to prevail, whereas for even number of such operators we face no problem since even number of anticommuting operators can be grouped to behave like bosonic operators. If we demand that all the states be even under  $(-1)^F$ , then half of states which had above mentioned undesirable feature, are removed including the tachyon. This is the GSO [40] projection. Moreover, after the unwanted states have been discarded from the spectrum, the remaining states of the theory belong to the representations of spacetime supersymmetry when we consider full spectrum of the superstring theory. Note that the operator  $(-1)^F$  is defined up to a sign ambiguity. If we choose the sign convention that the first excited state has  $(-1)^F = +1$  which arises due to action of  $\psi^i$ , on the ground state, then we can fix  $(-1)^F$  quantum number of the rest of the states. In this sign convention, tachyon will carry quantum number -1. There is another convention where tachyon has quantum number +1 and then the massless, first excited state, carries quantum number -1. The fermion numbers  $F_L$  and  $F_R$  can be introduced separately for the left and right moving sectors respectively. When one computes supercharge algebra with Ramond condition, the zero modes of the fermions in supercharge give an anomaly term besides the  $L_0$  term (that is Hamiltonian) and anomaly vanishes for  $D = 10$ . Moreover, the anticommutation relation of the R-zero modes are like Dirac gamma matrices carrying target space indices. One finds that massless states appear in the R-sector and they satisfy Dirac equation. They transform as ten dimensional spinors (S) or conjugate spinors (C). Since we are considering the left moving sector here at the moment, S has +1 eigenvalue and C has -1 eigenvalue under the  $(-1)^{F_L}$ . When we construct other excited states on these states they turn out to be massive. In view of this one need not apply GSO projection, no tachyon is to be removed.

When we combine the left and right moving sectors four combinations will appear in the description of the closed string spectrum. NS-NS, NS-R, R-NS and R-R; where the first sector is from left movers and second is from right movers in the above four combinations. Let us look at them one by one.

(i) NS-NS: The states are created due to the action of the creation operators from the left and the right moving sector. They will transform as tensors under 10-dimensional Lorentz transformation. After GSO projection is implemented, the lowest lying states are massless and they can be decomposed into three groups, symmetric traceless, antisymmetric tensor and a scalar under the rotations.

(ii) NS-R: The GSO projection, as discussed is  $(-1)^{F_L} = 1$  and one keeps the S representation of



the R sector here. The massless states consist of spacetime spinors.

(iii) R-NS: Here the GSO projection on NS sector from right side gives fermion number 1. We have the choice of keeping S spinor or the C spinor and obviously the states are spinorial.

(iv) R-R: The fermionic operators act from both sides and therefore, the resulting state will be bosonic in character. It will depend what combination we decide to keep. For example, if one keeps S from left side and  $\bar{C}$  from right side the product decomposes into a vector and a three index antisymmetric tensor (has to be antisymmetric - it arises from anticommuting objects). These belong to the bosonic sectors of type IIA theory. There is other combination which S from left and  $\bar{S}$  from right combine and their decomposition is a scalar, 2-form potential and 4-form (antisymmetric) potential whose field strengths are self-dual in ten dimensions and these states are bosonic sector of type IIB theory.

We are in a position to classify string theories according to their important characteristics. There are two 10-dimensional theories which have  $N = 2$  supersymmetry in target space. Their massless bosonic sectors are as follows: type IIA has graviton,  $G_{\mu\nu}$ , antisymmetric tensor  $B_{\mu\nu}$  and dilaton,  $\phi$ , coming from the NS-NS sector and a gauge potential  $A_\mu$  and three index antisymmetric tensor potential  $C_{\mu\nu\lambda}$ , coming from the R-R sector. These two theories have 32 generators of supersymmetry; type IIA is called non-chiral theory whereas type IIB is known as chiral theory. Although the bosonic fields coming from the RR sectors in these two string theories are tensors of different ranks, the total number of degrees of freedom of these tensors in each of the theories (A and B) are the same [41] and this can be checked by counting the physical degrees of freedom RR gauge fields of type IIA and IIB.

Next, we introduce the heterotic string which is very attractive when one tries to establish connection of string theory with the gauge groups of the standard model. The heterotic string, in ten dimensions, contains  $N = 1$  supergravity multiplet, super Yang-Mills gauge theory along with chiral fermions. There are two possible choices for the gauge groups:  $SO(32)$  or  $E_8 \times E_8$ , in the construction of the heterotic string. Therefore, heterotic string theory fulfills Green-Schwarz anomaly cancellation condition. Moreover, when the theory is compactified to four dimensions on Calabi-Yau manifold, the resulting theory has many features of the standard model and the gauge group  $SU(3) \times SU(2) \times U(1)$  can be embedded in the 4-dimensional theory. Let us briefly discuss how the heterotic string is constructed. We discussed the closed bosonic string and noted that the string coordinates can be decomposed to left movers and right movers and each can be expanded in Fourier modes. Moreover, the Virasoro generators are also separated into two groups, one group is expressed in terms of oscillators of one kind only (say left mover) and the other group of generators are expressed in terms of the oscillators of the other types (left movers). When one computes the quantum algebra, the anomaly free condition is imposed on each groups of Virasoro generators. In case of a closed string with worldsheet supersymmetry, same situation appears, because the fermion equations of motion is also written in terms of equations of motion of the Weyl Majorana fermions. If we were interested in constructing a string theory which satisfies requirements of conformal invariance, we could have a left moving closed bosonic string and a right moving superstring. The former will satisfy Virasoro algebra and latter the super Virasoro algebra. The triumph of the Heterotic string is that, when we look at the massless spectrum of the theory, it has  $N = 1$  spacetime supersymmetry, contains the appropriate gauge groups ( $SO(32)$  or  $E_8 \times E_8$ ) as is required for the consistency due to Green-Schwarz anomaly cancellation condition. Therefore, the closed bosonic string has 16 of its spatial coordinates compactified so that those coordinates themselves are periodic. Furthermore, using the standard techniques of  $1 + 1$  dimensional field theory, the compact bosonic coordinates could be fermionised to give 32 Weyl Majorana fermions which are left movers. Thus, we have 10 bosonic coordinates and their 10 super partners (in light-cone gauge 8 bosons and 8 fermions) in the right moving sector and 10 bosonic coordinates and 32 fermions (from compact coordinates) on the left moving sector. Whenever, we adopt NS boundary conditions for these fermions arising out of compactification, tachyon will appear in the spectrum. Of course, by introducing GSO projection on the right moving sector we shall have spacetime supersymmetry. So far as right moving part is concerned, bosonic states come from states with NS boundary condition and fermions arise due to the Ramond boundary conditions. The choice of boundary conditions on the left moving fermions (coming from compact directions),



give rise to two different types of gauge groups. (i) All the left moving fermions can satisfy R-type (periodic) boundary condition or they can satisfy NS-type boundary conditions. Then there is GSO condition which ensures that there are only states which have even number of these fermions (only one type boundary condition). Thus the massless bosonic spectrum is given by symmetric second rank tensor field, antisymmetric tensor field and a scalar together with 496 gauge bosons belonging to the adjoint representation of  $SO(32)$ . (ii) The second possible choice of boundary condition for the left moving fermions is to divide them to two groups containing 16 fermions. Now there are four choices of boundary conditions (a) All satisfy R boundary conditions, (b) periodic (R) boundary condition is imposed on both the groups, (c) all the fermions in first group (call it I) have R boundary condition and the group II has NS antiperiodicity and finally (d) group I belong to NS boundary condition and II are in R. The GSO projection is such that it keeps even number of fermions from each group in the spectrum in every sector. When one works out the bosonic spectrum, it contains again second rank symmetric tensor, antisymmetric tensor of rank two, the scalar, dilaton and 496 gauge bosons in the adjoint representation of  $E_8 \times E_8$ .

There is another superstring theory, known as type I. A simple way to describe type I string is from the perspective of IIB theory. Consider the parity operation  $\mathcal{P}$  on the worldsheet such that the 'spatial' coordinate  $\sigma \rightarrow -\sigma$  under  $\mathcal{P}$ . In type IIB theory,  $\mathcal{P}$  exchanges left and right moving sectors. Now, if we demand that we retain only those states which are invariant under  $\mathcal{P}$ , we get the type I string. In the NS-NS sector, graviton and the dilaton survive; the antisymmetric tensor is removed. From the RR sector, the only surviving field is the second rank antisymmetric tensor. Moreover, there are Weyl Majorana fermions and a gravitino surviving the operation giving rise to  $N = 1$  supergravity multiplet. The open string states are also included in type I spectrum. In this case, the worldsheet degrees of freedom are same as in the closed string case. One imposes Neumann boundary conditions on the bosonic coordinates and suitable boundary conditions on worldsheet fermions. The gauge group that can get attached to the open string is  $SO(32)$  and thus there is corresponding super Yang-Mills theory besides the states we mentioned above.

Thus there are five perturbatively consistent string theories. The scattering of particles belonging to spectrum of a string theory can be described by introducing vertex operators [43]. They are required to satisfy constraints due to conformal or superconformal transformations. They must transform as representations of Lorentz group, like a wave function. In the first quantized frame work, one can calculate scattering of these particles in a well defined perturbation theory. It is one of the great virtues of the superstring theories that all these calculations are ultraviolet finite. Therefore, we have five different string theories in ten dimensions.

One of the most efficient ways to study properties of string theory is to investigate the evolution of a string in the background of its massless excitations and then explore the consequences of conformal invariance for such a situation. Let us consider closed bosonic string in the background of its massless excitations such as graviton, antisymmetric tensor and dilaton. The action (2.3) generalizes to

$$-\frac{T}{2} \left( \int d^2\sigma \{ \sqrt{-\gamma} \gamma^{ab} G_{\mu\nu}(X) + \epsilon^{ab} B_{\mu\nu}(X) \partial_a X^\mu \partial_b X^\nu \} + \frac{1}{2} \int d^2\sigma \sqrt{-\gamma} R^{(2)} \phi(X) \right) \quad (2.32)$$

Here  $R^{(2)}$  is the scalar curvature of the worldsheet computed with  $\gamma_{ab}$ . The first two terms show the couplings of  $G_{\mu\nu}$  and  $B_{\mu\nu}$  to the string coordinates. In close string theory there is a massless state which transforms as symmetric second rank tensor and it is identified as graviton and there is an antisymmetric massless second rank tensor state. The above action describes motion of the string in the background of these massless states,  $G_{\mu\nu}$  and  $B_{\mu\nu}$ ; the last term is the coupling of the string to the massless scalar, the dilaton. This is an action for a two dimensional  $\sigma$ -model and we can interpret that  $G_{\mu\nu}$  and  $B_{\mu\nu}$  play the role of coupling constants. At the classical level the dilaton coupling breaks the conformal invariance explicitly. However, it is important to explore the consequences of the quantum invariance as we have seen that the quantum invariance principle imposes strong constraints on the theory. There is a well defined procedure to compute the conformal anomaly for such theories[44]. One of the ways to ensure conformal invariance of the quantum theory is to demand that the two dimensional energy momentum stress tensor has vanishing trace. As is well known, the conformal anomaly is related to the corresponding  $\beta$ -function

of the theory. Thus, vanishing of the  $\beta$ -functions will ensure conformal invariance. Moreover, the beta functions can be computed order by order in the  $\sigma$ -model perturbation theory;  $\alpha'$  being the expansion parameter. The relevant  $\beta$ -functions are:

$$\frac{\beta_{\mu\nu}^G}{\alpha'} = R_{\mu\nu} - \frac{1}{4}H_{\mu\rho\lambda}H_{\nu}^{\rho\lambda} + \nabla_{\mu}\nabla_{\nu}\phi \quad (2.33)$$

$$\frac{\beta_{\mu\nu}^B}{\alpha'} = \nabla^{\rho}[e^{-\phi}H_{\mu\nu\rho}] \quad (2.34)$$

$$\beta^{\phi} = \Lambda + 3\alpha'[(\nabla\phi)^2 - 2\nabla^{\mu}\nabla_{\mu}\phi - R + \frac{1}{12}H^2] \quad (2.35)$$

The notations are as follows:  $R_{\mu\nu}$  is the Ricci tensor for the target space computed from the string frame metric  $G_{\mu\nu}$ .  $\Lambda = D - 26$  or  $D - 10$  depending on whether we are dealing with a pure bosonic string or superstring (if we deal with superstring the coupling of worldsheet fermions to the background has to be taken into account),  $D$  being the spacetime dimension.  $H_{\mu\nu\rho} = \partial_{\mu}B_{\nu\rho} + \text{cycl.perm.}$ , is the field strength of two form potential  $B_{\mu\nu}$ . It might be worthwhile to point out that for the constant value of dilaton the last term in (2.32) is just the Euler character of the surface. When we write the path integral form with the action, we see that the factor  $e^{-\chi\phi_0}$  comes out; where  $\chi$  is the Euler character and  $\phi_0$  is the constant value of the dilaton. In this light the string coupling constant is defined as

$$g_{str} = e^{\phi_0/2} \quad (2.36)$$

Let us look for an action in the target space such that the variation of that action with respect to the backgrounds  $G_{\mu\nu}$ ,  $B_{\mu\nu}$  and  $\phi$  would reproduce the  $\beta$ -function equations we have obtained earlier. We also know that these  $\beta$ -functions must vanish (to the order in  $\alpha'$  they are computed) in order to respect conformal invariance of the theory.

The resulting action is

$$S = \int d^Dx \sqrt{-G} e^{-\phi} [R + (\partial\phi)^2 - \frac{1}{12}H^2] \quad (2.37)$$

This action is called the tree level string effective action. Solutions of the equation of motion of this action (same as solution to  $\beta$ -function equation) correspond to admissible background configurations with respect conformal invariance. In other words, every solution is an acceptable vacuum of the string theory to lowest order in  $\alpha'$  since the effective action is obtained from the  $\beta$ -function equations keeping only lowest order terms in  $\sigma$ -model perturbation theory. Therefore, if we find solutions which correspond to cosmological situation with given  $G$ ,  $B$  and  $\phi$ , or a black hole solution, or a wormhole solution all these types of geometries with the appropriate matter content, consistent with the equations of motion, can be interpreted as string vacuum backgrounds.

So far we have been discussing the quantization of string theories and examining the consequences of conformal invariance. Note that all the consistent string theories are defined in spacetime dimensions higher than four i.e.  $D = 10$ . Therefore, one must answer the question what these theories have to do with the spacetime where we live. This issue has been taken up by Kaluza and Klein more than seven decades ago. The basic idea is rather simple. In order to construct a unified theory of gravity and electrodynamics, they considered an Einstein-Hilbert type action in 5-spacetime dimensions which is invariant under general coordinate transformations in five dimensions. Let us imagine that one of the dimensions, the 5th one, is a circle of very small radius which could not be probed today using any particle whose de Broglie wave length is comparable to the size of that circle. Then we shall not be aware of this scale. Let us assume, to a first approximation, that the metric does not depend on the 5th coordinate. Kaluza and Klein showed that the resulting theory looks like Einstein theory and Maxwell theory in four dimensions. What was general coordinate invariance in 5-dimensional theory, turned out to be general coordinate transformation and Abelian gauge transformation (of Maxwell theory) in four dimensions. Although, the original Kaluza-Klein proposal had many shortcomings, the idea is very relevant for construction of four dimensional theories starting from the 10-dimensional string theories in the present context. We shall explore this aspect and we shall see how duality symmetries arise for compactified string theories. We shall set  $T = 1$  from now on, whenever, we shall need to introduce the slope parameter/tension, we shall explicitly mention in that context.

### 3 Duality Symmetries in String Theory

One of the marvels of the string theory is its rich symmetry structure. We have noticed how the conformal invariance imposes strong constraints on the theory: when we consider flat spacetime the dimensionality is fixed by this symmetry. On the other hand if we consider strings in backgrounds, we get the equations of motion for them by demanding that the corresponding  $\beta$ -functions must vanish. Moreover, there are local symmetries like invariance associated with general coordinate transformation due to the presence of the graviton and an Abelian gauge symmetry since the antisymmetric tensor is also a part of the massless multiplet.

The duality symmetries play a crucial role in understanding various features of string theory. Since string is an extended object, there are symmetries special to string theory. Consider a particle whose motion is on a circular path, the momentum is quantized in suitable units of the inverse radius in order that the wave function maintains single valuedness. However, in case of a string, one of whose coordinate has geometry of a circle, offers more interesting possibilities. In fact a string theory with one spatial direction compactified as  $S^1$  of radius  $R$  cannot be distinguished from another theory whose coordinate is compactified on a circle of radius  $\frac{1}{R}$ . Let the compactified coordinate be denoted by  $Y(\sigma, \tau)$  with the periodicity condition

$$Y(\sigma, \tau) + 2\pi R = Y(\sigma, \tau) \quad (3.1)$$

Furthermore, the string coordinate is also periodic when  $\sigma$  goes over  $2\pi$  for the closed string. Since, the coordinate is compact, zero momentum mode must be quantized to maintain single valuedness of the wave function just as the case in field theory. In case of the string, the string can wind around the compact direction. It will cost more energy if the string winds  $m$ -number time, because it will have to stretch more. Therefore, the effect due to windings has to be taken into account too while estimating energy levels [45]. Thus the mode expansions for left and right moving sectors are:

$$Y_R = y_R + \sqrt{\frac{1}{2}} p_R (\tau - \sigma) + \text{oscillators} \quad (3.2)$$

$$Y_L = y_L + \sqrt{\frac{1}{2}} p_L (\tau + \sigma) + \text{oscillators} \quad (3.3)$$

The momentum zero modes  $p_{R,L}$  will have the following form to be consistent with what we said earlier

$$p_R = \frac{1}{\sqrt{2}} \left( \frac{n}{R} - Rm \right), \text{ and } p_L = \frac{1}{\sqrt{2}} \left( \frac{n}{R} + Rm \right) \quad (3.4)$$

The above equation states that in general the contribution of the Kaluza-Klein mode is  $\frac{1}{R}$  times an integer and the winding mode is an integer times the radius. The total momentum is just  $P = \frac{1}{\sqrt{2}}(p_R + p_L)$ , which is integral of momentum density over  $\sigma$ . The total Hamiltonian is

$$H = L_0 + \bar{L}_0 = \frac{1}{2}(p_L^2 + p_R^2) + \text{oscillators} \quad (3.5)$$

Now we consider the general case of toroidal compactification and present the derivation as was done in reference [49]. Let  $G_{\alpha\beta}$  and  $B_{\alpha\beta}$  be constant backgrounds,  $\alpha, \beta = 1, \dots, d$ , and  $Y^\alpha(\sigma, \tau)$  are the string coordinates. The two-dimensional  $\sigma$ -model action containing these coordinates is

$$I_{compact} = \frac{1}{2} \int d^2\sigma [G_{\alpha\beta} \eta^{ab} \partial_a Y^\alpha \partial_b Y^\beta + \epsilon^{ab} B_{\alpha\beta} \partial_a Y^\alpha \partial_b Y^\beta] \quad (3.6)$$

where  $G_{\alpha\beta}$  and  $B_{\alpha\beta}$  are constant backgrounds. The coordinates are taken to satisfy the periodicity conditions  $Y^\alpha \simeq Y^\alpha + 2\pi$ . Here we take the compactification radius to be unity for simplicity in calculations. For closed strings it is necessary that

$$Y^\alpha(2\pi, \tau) = Y^\alpha(0, \tau) + 2\pi m^\alpha \quad (3.7)$$

where the integers  $m^\alpha$  are called winding numbers. It follows from the single-valuedness of the wave function on the torus that the zero modes of the canonical momentum,  $P_\alpha = G_{\alpha\beta}\partial_\tau Y^\beta + B_{\alpha\beta}\partial_\sigma Y^\beta$ , are also integers  $n_\alpha$ . Therefore the zero modes of  $Y^\alpha$  are given by

$$Y_0^\alpha = y^\alpha + m^\alpha \sigma + G^{\alpha\beta}(n_\beta - B_{\beta\gamma}n^\gamma)\tau \quad (3.8)$$

where  $G^{\alpha\beta}$  is the inverse of  $G_{\alpha\beta}$ . The Hamiltonian is given by

$$\mathcal{H} = \frac{1}{2}G_{\alpha\beta}(\dot{Y}^\alpha \dot{Y}^\beta + Y'^\alpha Y'^\beta) \quad (3.9)$$

where  $\dot{Y}^\alpha$  and  $Y'^\beta$  are derivatives with respect to  $\tau$  and  $\sigma$ , respectively. Let us elaborate a little bit on the significance of what we have done with respect to the compact coordinates. Since the coordinates  $Y^\alpha$ , are compact, they satisfy eq.(3.7). Moreover, these coordinates can be expanded as usual in terms of their zero modes and the oscillators. However, for the discussion of T-duality, we focus our attentions on the zero mode parts and the contribution of these parts to the Hamiltonian, given above.

Since  $Y^\alpha(\sigma, \tau)$  satisfies the free wave equation, we can decompose it as the sum of left- and right-moving pieces. The zero mode of  $P^\alpha = G^{\alpha\beta}P_\beta$  is given by  $p_L^\alpha + p_R^\alpha$  where

$$p_L^\alpha = \frac{1}{2}[m^\alpha + G^{\alpha\beta}(n_\beta - B_{\beta\gamma}m^\gamma)] \quad (3.10)$$

$$p_R^\alpha = \frac{1}{2}[-m^\alpha + G^{\alpha\beta}(n_\beta - B_{\beta\gamma}m^\gamma)] \quad (3.11)$$

The mass-squared operator, which corresponds to the zero mode of  $\mathcal{H}$ , is given (aside from a constant) by

$$(mass)^2 = G_{\alpha\beta}(p_L^\alpha p_L^\beta + p_R^\alpha p_R^\beta) + \sum_{m=1}^{\infty} \sum_{i=1}^d (\alpha_{-m}^i \alpha_m^i + \bar{\alpha}_{-m}^i \bar{\alpha}_m^i) \quad (3.12)$$

As usual,  $\{\alpha_m\}$  and  $\{\bar{\alpha}_m\}$  denote oscillators associated with right- and left-moving coordinates, respectively. Substituting the expressions for  $p_L$  and  $p_R$ , the mass squared can be rewritten as

$$(mass)^2 = \frac{1}{2}G_{\alpha\beta}m^\alpha m^\beta + \frac{1}{2}G^{\alpha\beta}(n_\alpha - B_{\alpha\gamma}m^\gamma)(n_\beta - B_{\beta\delta}m^\delta) + \sum (\alpha_{-m}^i \alpha_m^i + \bar{\alpha}_{-m}^i \bar{\alpha}_m^i) \quad (3.13)$$

It is significant that the zero mode portion of (3.13) can be expressed in the form

$$(M_0)^2 = \frac{1}{2}(m \ n)M^{-1} \begin{pmatrix} m \\ n \end{pmatrix}, \quad (3.14)$$

where  $M$  is the  $2d \times 2d$  symmetric matrix expressed in terms of constant backgrounds  $G$  and  $B$

$$M = \begin{pmatrix} G^{-1} & -G^{-1}B \\ BG^{-1} & G - BG^{-1}B \end{pmatrix} \quad (3.15)$$

In order to satisfy  $\sigma$ -translation symmetry, the contributions of left- and right-moving sectors to the mass squared must agree;  $L_0 = \bar{L}_0$ . The zero mode contribution to their difference is

$$G_{\alpha\beta}(p_L^\alpha p_L^\beta - p_R^\alpha p_R^\beta) = m^\alpha n_\alpha \quad (3.16)$$

Since this is an integer, it always can be compensated by oscillator contributions, which are also integers.

Equation (3.16) is invariant under interchange of the winding numbers  $m^\alpha$  and the discrete momenta  $n_\alpha$ . Indeed, the entire spectrum remains invariant if we interchange  $m^\alpha \leftrightarrow n_\alpha$  simultaneously let [46]

$$(G - BG^{-1}B) \leftrightarrow G^{-1} \quad \text{and} \quad BG^{-1} \leftrightarrow -G^{-1}B$$

These interchanges precisely correspond to inverting the  $2d \times 2d$  matrix  $M$ . This is the duality transformation generalizing the well-known duality  $R \leftrightarrow \frac{1}{R}$  in the  $d = 1$  case earlier. The general duality symmetry implies that the  $2d$ -dimensional Lorentzian lattice by the vectors  $\sqrt{2}(p_L^\alpha, p_R^\alpha)$  with inner product

$$\sqrt{2} (p_L, p_R) \cdot \sqrt{2} (p'_L, p'_R) \equiv 2G_{\alpha\beta}(p_L^\alpha p'_L{}^\beta - p_R^\alpha p'_R{}^\beta) = (m^\alpha n'_\alpha + m'^\alpha n_\alpha)$$

is even and self-dual ([47]). For toroidally compactified string theory, the coordinate periodicity condition and the conjugate momenta belong to the dual space and are suitable units. Furthermore, one can define corresponding metric to introduce the coordinates and their dual momentum vectors and define an inner product also. On the lattices the space of the coordinates (since the coordinates satisfy periodicity conditions) is the same as the dual space, then the lattice is called self-dual. Of special interest are the spaces where the length of the vector is even (with the definition of norm) we have even self-dual lattice. These types of lattices are very important in consistent theories with nonabelian gauge groups and to satisfy consistency requirements of

The moduli space parametrized by  $G_{\alpha\beta}$  and  $B_{\alpha\beta}$  is locally the coset  $O(d, d)/O(d, d)$ . The global geometry requires also modding out the group of discrete symmetries  $B_{\alpha\beta} \rightarrow B_{\alpha\beta} + N_{\alpha\beta}$  and  $G + B \rightarrow (G + B)^{-1}$ . These symmetries generate the  $C$  of  $O(d, d)$ . An  $O(d, d, Z)$  transformation is given by a  $2d \times 2d$  matrix  $A$  having and satisfying  $A^T \eta A = \eta$ , where  $\eta$  consists of off-diagonal unit matrices defining  $O(d, d, Z)$  transformation

$$\begin{pmatrix} m \\ n \end{pmatrix} \rightarrow \begin{pmatrix} m' \\ n' \end{pmatrix} = A \begin{pmatrix} m \\ n \end{pmatrix} \quad \text{and} \quad M \rightarrow AMA^T$$

It is evident that

$$m \cdot n = \frac{1}{2} \begin{pmatrix} m & n \end{pmatrix} \eta \begin{pmatrix} m \\ n \end{pmatrix}$$

$$\eta = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

which appears in eq.(3.16), and  $M_0^2$  in eq.(3.13) are preserved under the transformation that  $\eta$  is symmetric  $2d \times 2d$  matrix with off diagonal elements which are discrete. The crucial fact, already evident from the spectrum, is that toroidally compactified string theory certainly does not share the full  $O(d, d)$  symmetry of the low energy effective theory invariant under the discrete  $O(d, d, Z)$  subgroup.

So far, in discussing issues of compactifications, we have considered situations where all coordinates are compact. However, one can envisage the scenario, when some of the coordinates are compactified and the rest are noncompact. Furthermore, we treat the background fields as constant; however, in more realistic situations the backgrounds show noncompact coordinates. This is the more interesting situation where the string theory and six of its spatial coordinates are compactified on a circle, while the other six are noncompact. We shall add to the theory a dimensional reduction [48, 49, 50] so that we can compactify an arbitrary number of dimensions, so that the effective theory is defined in a lower spacetime dimension. This is useful, since the duality conjectures are in various spacetime dimensions related by the web of dualities in diverse dimensions.

The starting point is to consider the string effective action in  $D$  dimensions, coordinates, metric and all other tensors in the  $\hat{D}$  dimensional spacetime. The coordinates in  $D$ -dimensional spacetime are denoted by  $x^\mu$ . Therefore,  $\hat{D} = D + d$ . The theory is compactified on a  $d$ -dimensional

spacetime. The coordinates on the torus, sometimes referred to coordinates of internal dimensions, are denoted as  $y^\alpha, \alpha = 1, \dots, d$ . The bosonic part of the action is given by

$$\hat{S} = \int d^D x \sqrt{-\hat{G}} e^{-\hat{\phi}} \left[ \hat{R}(\hat{G}) + \hat{G}^{\mu\nu} \partial_\mu \hat{\phi} \partial_\nu \hat{\phi} - \frac{1}{12} \hat{H}_{\hat{\mu}\hat{\nu}\hat{\rho}} \hat{H}^{\hat{\mu}\hat{\nu}\hat{\rho}} \right]. \quad (3.22)$$

Note that  $\hat{S}$  is the bosonic part of the string effective action with backgrounds coming from NS-NS sector.  $\hat{H}$  is the field strength of antisymmetric tensor and  $\hat{\phi}$  is the dilaton. The backgrounds are taken to be independent of the internal coordinates,  $y^\alpha$  of the torus. Consequently, any transformations of the coordinated  $y^\alpha, \alpha = 1, 2, \dots, d$  does not affect the background fields and we recognize that there are  $d$  isometries. Furthermore, associated with these isometries, there will be  $d$  Abelian gauge fields since the  $\hat{D}$ -dimensional metric will have components carrying a  $D$ -dimensional spacetime index and an internal index  $\alpha$ . There will be components of the  $\hat{D}$ -dimensional metric which will carry indices of the toroidal coordinates, say  $\alpha, \beta$  and these will transform as scalars, often refer to as moduli. Similarly, if we consider the components of the  $\hat{D}$ -dimensional antisymmetric tensor field it will have  $D \times D$  component antisymmetric tensor,  $d$  Abelian gauge fields coming from spacetime and internal component and  $d \times d$  dimensional moduli (antisymmetric) when considered from  $D$ -dimensional point of view.

The metric  $\hat{G}_{\hat{\mu}\hat{\nu}}$  can be decomposed as

$$\hat{G}_{\hat{\mu}\hat{\nu}} = \begin{pmatrix} \mathcal{G}_{\mu\nu} + A_\mu^{(1)\gamma} A_\nu^{(1)\gamma} & A_{\mu\beta}^{(1)} \\ A_{\nu\alpha}^{(1)} & G_{\alpha\beta} \end{pmatrix}, \quad (3.23)$$

where  $G_{\alpha\beta}$  is the internal metric and  $\mathcal{G}_{\mu\nu}$ , the  $D$ -dimensional space-time metric, depend on the coordinates  $x^\mu$ . Note the appearance of Abelian gauge fields  $A^{(1)\alpha}$  due to the presence of the isometries. We also expect same number of gauge fields from the antisymmetric tensor  $\hat{B}_{\hat{\mu}\hat{\nu}}$ . Thus The dimensionally reduced action is,

$$\begin{aligned} S_D = \int d^D x \sqrt{-g} e^{-\phi} \left\{ R + g^{\mu\nu} \partial_\mu \phi \partial_\nu \phi - \frac{1}{12} H_{\mu\nu\rho} H^{\mu\nu\rho} \right. \\ \left. + \frac{1}{8} \text{tr}(\partial_\mu M^{-1} \partial^\mu M) - \frac{1}{4} \mathcal{F}_{\mu\nu}^i (M^{-1})_{ij} \mathcal{F}^{\mu\nu j} \right\}. \end{aligned} \quad (3.24)$$

Here  $\phi = \hat{\phi} - \frac{1}{2} \log \det G$  is the shifted dilaton.

$$H_{\mu\nu\rho} = \partial_\mu B_{\nu\rho} - \frac{1}{2} A_\mu^i \eta_{ij} \mathcal{F}_{\nu\rho}^j + (\text{cyc. perms.}), \quad (3.25)$$

$\mathcal{F}_{\mu\nu}^i$  is the  $2d$ -component vector of field strengths

$$\mathcal{F}_{\mu\nu}^i = \begin{pmatrix} F_{\mu\nu}^{(1)\alpha} \\ F_{\mu\nu}^{(2)} \end{pmatrix} = \partial_\mu \mathcal{A}_\nu^i - \partial_\nu \mathcal{A}_\mu^i, \quad (3.26)$$

$A_{\mu\alpha}^{(2)} = \hat{B}_{\mu\alpha} + B_{\alpha\beta} A_\mu^{(1)\beta}$  (recall  $B_{\alpha\beta} = \hat{B}_{\alpha\beta}$ ), and the  $2d \times 2d$  matrices  $M$  and  $\eta$  are defined as

$$M = \begin{pmatrix} G^{-1} & -G^{-1}B \\ BG^{-1} & G - BG^{-1}B \end{pmatrix}, \quad \eta = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \quad (3.27)$$

Note that the elements of the matrix  $M$ ,  $G_{\alpha\beta}$  and  $B_{\alpha\beta}$  depend on spacetime coordinates  $x^\mu$  in contrast to the earlier case (3.15) where those back grounds were taken to be constant. The action (3) is invariant under a global  $O(d, d)$  transformation,

$$M \rightarrow \Omega^T M \Omega, \quad \Omega \eta \Omega^T = \eta, \quad \mathcal{A}_\mu^i \rightarrow \Omega^i_j \mathcal{A}_\mu^j, \quad \text{where } \Omega \in O(d, d). \quad (3.28)$$

and the shifted dilaton,  $\phi$ , remains invariant under the  $O(d, d)$  transformations. Moreover,  $M \in O(d, d)$  and  $M^T \eta M = \eta$ . Thus if we solve for a set of backgrounds,  $M, \mathcal{F}$  and  $\phi$ , satisfying the equations of motion they correspond to a vacuum configuration of the string theory. The  $O(d, d)$  symmetry is known as the target space duality (or T-duality) symmetry, it is a stringy symmetry and there is no analogue of winding modes in ordinary field theory. The symmetry holds good order by order in string perturbation theory. Therefore, predictions of T-duality can be tested within the frame work of perturbation theory. We remark in passing that, if we had considered an effective action in  $\hat{D}$  dimensions with  $n$  Abelian gauge fields, the reduced action in  $D$  dimensions will be invariant under  $O(d, d + n)$  symmetry. This is of importance, since in case of the heterotic string, the ten dimensional action with 16 Abelian gauge fields corresponding to the Cartan subalgebra of the nonabelian gauge groups of the theory, when reduced to lower dimensions will exhibit the symmetry  $O(d, d + n)$  we mentioned.

Thus if we have a set of background configurations it is possible to generate another set of gauge inequivalent backgrounds by implementing suitable  $O(d, d)$  transformations. The new backgrounds will also satisfy the equations of motion and they will be acceptable vacuum configurations. In fact the  $O(d, d)$  symmetry was discovered for non-constant backgrounds in the context of cosmological solutions in string theory [51, 52], when the backgrounds carried only time dependence. One could generate new cosmological solutions through  $O(d, d)$  transformations [53, 54]. The applications of  $O(d, d)$  transformations in the context of black holes was to generate new black hole solutions was initiated by Sen [55] and there is a vast literature in this subject [19, 31].

Next, we discuss S-duality in string theory. This symmetry relates a theory in the weak coupling regime to a theory in the strong coupling domain. In some it is the same theory which gets related to itself, like the type IIB theory. In some other situations one theory gets related to another one: a familiar example is that heterotic string compactified on  $T^4$  is related to type IIA theory compactified on  $K_3$ . A simple example is the Maxwell electrodynamics. The equations are invariant under  $\mathbf{E} \rightarrow \mathbf{B}$  and  $\mathbf{B} \rightarrow -\mathbf{E}$ . However, in the presence of sources, one has to be careful. The usual Maxwell equations have only sources carrying electric charges and then the equations are not symmetric under the above duality transformations. Thus it is necessary to add sources carrying magnetic charges to maintain electric-magnetic duality. This led Dirac to formulate the theory of magnetic monopoles. As is well known, the existence of magnetic monopole in the theory leads to the famous charge quantization condition:  $e \cdot g = 2\pi n$ , where  $e$  is the electric charge and  $g$  is the magnetic charge. This relation has profound implications; if the theory of electrically charged particles is described by a small coupling constant (indeed fine structure constant  $\alpha = \frac{1}{137}$ ), then the theory describing magnetic monopoles will have large value for such charges corresponding to strong coupling constant. In the case of gauge theories with spontaneous symmetry breaking, magnetic monopoles appear as classical solutions of nonlinear field equations [56, 57]. Note that the electric charge in such theories are obtained from the Noether currents whereas, the magnetic charge of 't Hooft-Polyakov monopoles are of topological nature. The charges respect the Dirac quantization condition. Furthermore, the massive gauge bosons (acquiring mass through Higg's mechanism) have masses proportional to the gauge coupling constant, whereas the monopole masses are inversely proportional to the gauge coupling constant (electric charge). Consequently, if the gauge bosons are light in a SSB theory, the monopoles are heavy; indeed the monopoles have the interpretation of being the solitons of the theory. One of the most fundamental contributions to developments in S-duality came from the work of Montonen and Olive [58]. According to them, we might envisage a dual formulation of fundamental physics where the role of Noether charges and topological charges are interchanged. One can visualise that monopoles will appear as elementary particles and the W-bosons will be solitonic counter parts. In fact one could check their mass formula  $m^2 = C(e^2 + g^2)$ ; where  $C$  is related to VEV of Higgs in SSB theories. In fact W boson and photon satisfy this formula. If a particle had been discovered with magnetic charge this relation could be verified. Since it is symmetric under the interchange of  $e$  and  $g$  and Dirac's rule tells us that  $e$  and  $g$  are related, one could formulate the theory in the dual picture. However, the monopole mass obtained in SSB theory is a classical one and it is subject to quantum corrections. Thus, Montonen-Olive idea could not be consistently checked in usual field theories. There are special types of supersymmetric field theories where there is no quantum correction to the mass

and furthermore, the W-bosons and monopoles belong to the same multiplet. In such cases there is the possibility of checking this conjecture.

We recall that the Yang-Mills theory also admits the introduction of the  $\theta$  term in its action. Thus, gauge theories have two parameters, the Yang-Mills coupling constant  $e$  and the  $\theta$  parameter. The latter couples to the field strengths as follows:

$$-\frac{\theta e^2}{32\pi^2} F_{\mu\nu}^a \tilde{F}_a^{\mu\nu}, \quad (3.29)$$

where  $\tilde{F}^{a\mu\nu} = \epsilon^{\rho\lambda}_{\mu\nu} F_{\rho\lambda}^a$ . Note that this term is a surface term and does not contribute to classical equations of motion and presence of this term does not affect renormalizability in the perturbation theory. It was noted by Witten [59] that in the presence of monopoles, this term shifts the allowed values of the electric charge in the monopole sector. Thus we can have electrically charged, magnetically charged particles and a third kind of particles carrying both the charges. The Yang-Mills Lagrangian can be written in the following form after taking into account the effect of the  $\theta$  term and introducing a complex coupling constant  $\tau = \frac{\theta}{2\pi} + \frac{4i\pi}{e^2}$

$$\mathcal{L} = -\frac{1}{32\pi} \text{Im}(\tau [F_a^{\mu\nu} + i\tilde{F}_a^{\mu\nu}][F_{\mu\nu}^a + i\tilde{F}_{\mu\nu}^a]) \quad (3.30)$$

Following qualitative argument tells us about the strong-weak duality group. (i) When  $\theta$  goes over its period  $2\pi$  physics is the same. Thus, we expect that the theory be invariant when  $\tau \rightarrow \tau + 1$ . (ii) We also know that, under electric magnetic duality,  $\tau \rightarrow -\frac{1}{\tau}$ . One can argue that, when  $\theta$  is arbitrary, the duality group is generated by these transformations. Thus, the duality group is identified to be  $SL(2, Z)$ . Therefore, in a theory with  $SL(2, Z)$  symmetry one could check the spectrum with charged particles, monopoles and dyons. The complex coupling constant  $\tau$  is often referred to as modular parameter or moduli. Moreover, when we discuss strong-weak duality in the context of string theory, dilaton and axion will be combined to define the moduli field. As mentioned earlier, string theory does not admit any arbitrary parameters as coupling constants. All the coupling constants appear as VEV of some scalar fields, i.e. moduli. Therefore, very often, the term coupling constants and moduli are used interchangeably in string theory.

As mentioned earlier, the mass formulas are protected from quantum corrections in supersymmetric theory. Moreover, some of the solitonic solutions in the supersymmetric theories satisfy special properties: (i) They saturate the BPS bound and (ii) these solutions preserve a part of the supersymmetry of the original theory. These attributes play a very important part in testing duality conjectures in field theory and in string theory. In order to illustrate the basic point, let us consider a two dimensional example due to Witten and Olive [60], where the field content is a scalar field and Majorana fermion. The Lagrangian density is

$$\mathcal{L} = \frac{1}{2}[(\partial_\mu \Phi)^2 + i\bar{\Psi}\gamma^\mu \partial_\mu \Psi - V^2(\Phi) - V'(\Phi)\bar{\Psi}\Psi] \quad (3.31)$$

The potential is arbitrary function of  $\Phi$  and 'prime' denotes derivative with respect to  $\Phi$ . As was the case in worldsheet supersymmetry, we can work in terms of chiral components of fermions and the two super charges are

$$Q_+ = \int dx[(\partial_0 + \partial_1)\Phi\Psi_+ - V(\Phi)\Psi_-] \quad (3.32)$$

$$Q_- = \int dx[(\partial_0 - \partial_1)\Phi\Psi_- + V(\Phi)\Psi_+] \quad (3.33)$$

In light-cone variables  $Q_\pm^2 = P_\pm$ , with  $P_\pm = P_0 \pm P_1$  and it turns out that  $\{Q_+, Q_-\} = 0$ , in most of the case. However, careful analysis shows that the anticommutator, is proportional to a surface integral

$$\{Q_+, Q_-\} = 2 \int dx \frac{\partial}{\partial x} H(\Phi) \quad (3.34)$$



and  $H'(\Phi) = V(\Phi)$ . This surface integral does not necessarily vanish when one considers solitonic states. If we denote the R.H.S. of (3.34) by the operator  $T$ , then it can be evaluated for the case at hand. Now the algebra of charges are different from usual case and one can write

$$P_+ + P_- = T + (Q_+ - Q_-)^2 \quad (3.35)$$

$$P_+ + P_- = -T + (Q_+ + Q_-)^2 \quad (3.36)$$

The R.H.S. of each equation above has a piece which is a complete square and we have  $P_+ + P_- \geq |T|$ . If we consider single particle of mass  $M$  and go to its rest frame  $P_\pm = M$ ; we arrive at

$$M \geq |T| \quad (3.37)$$

The bound will be equality when we have states,  $|s\rangle$  such that  $(Q_+ + Q_-)|s\rangle = 0$  or  $(Q_+ - Q_-)|s\rangle = 0$ . The bound on  $M$  is the Bogomolny bound. The state which saturates it is called a BPS state. This bound also can be derived in a Lorentz covariant manner. We note that, for the states saturating the BPS bound, only half of the supersymmetries are preserved. In string theory or field theories with large number of supersymmetries, the algebra of the charges for a set of charges  $\{Q_\alpha\}$ ,  $\alpha = 1, \dots, N$ , can be brought to the form

$$\{Q_\alpha, Q_\beta\} = \delta_{\alpha\beta} \quad (3.38)$$

This will be possible if there are no states which are annihilated by some of these charges and in that case, we shall get supermultiplets as usual. However, just like the soliton case considered earlier, if there are states which will be annihilated by some charges then we shall have a situation where

$$\{Q_a, Q_b\} = \delta_{ab}, \quad \text{for } a, b = 1, \dots, M \quad (3.39)$$

$$\{Q_\alpha, Q_\beta\} = 0, \quad \alpha, \beta = M + 1, \dots, N \quad (3.40)$$

So we see that these states will be lower dimensional representations since  $M < N$ . Again, citing the example of two dimensional case, we can state the general result that when there are soliton like states getting annihilated by some of the supercharges, then the symmetric matrix  $\{Q_\alpha, Q_\beta\}$  will have some zero eigen values. The charges (analog of  $T$ ) and masses get related in the process. This is true for monopoles in 4-dimensional theories. The string effective action is defined in 10 dimensions and one can seek solutions for extended objects in space and there are BPS states in this regime too.

Let us compactify the heterotic string effective action on  $T^6$  to come to a four dimensional theory. As mentioned earlier, the T-duality group is  $O(6, 22)$  with scalars parametrizing the moduli  $\frac{O(6, 22)}{O(6) \times O(22)}$ , 28 gauge bosons, graviton  $G_{\mu\nu}$  and antisymmetric tensor  $B_{\mu\nu}$ . The four dimensional effective action for the heterotic string, following the prescriptions of [49], can be obtained in a straight forward manner. The T-duality invariance is manifest when we are in the string frame metric with shifted dilaton  $\hat{\phi} - \frac{1}{2} \ln \det G_{\alpha\beta}$ . However, when one considers the S-duality properties of the theory, it is convenient to go over to the Einstein frame metric,  $g_{\mu\nu}$  through the conformal transformation,  $g_{\mu\nu} = e^{-\phi} G_{\mu\nu}$ . In string theory, all the coupling constants are related to the VEV of the dilaton and therefore, in order to identify the parameters of S-duality group, we have to choose the field whose VEV will coincide with the  $\theta$  parameter. Notice that the field strength of antisymmetric tensor,  $H_{\mu\nu\rho}$  has only one degree of freedom in four dimensions when we fix all gauge freedoms. In fact, if we take dual of this field, it is a pseudoscalar particle and that is what we need, an axion. The starting point is the four dimensional effective action [64] with Einstein frame spacetime metric

$$S^{(4)} = \int_M dx \sqrt{-g} \left\{ R - \frac{1}{2} g^{\mu\nu} \partial_\mu \phi \partial_\nu \phi + \mathcal{L}_2 + e^{-\phi} \mathcal{L}_3 + e^{-2\phi} \mathcal{L}_4 \right\} \quad (3.41)$$

with  $\mathcal{L}_2$ ,  $\mathcal{L}_3$ , and  $\mathcal{L}_4$  defined as follows

$$\mathcal{L}_2 = \frac{1}{8} \text{tr}(\partial_\mu M^{-1} \partial^\mu M) . \quad (3.42)$$

$$\mathcal{L}_3 = -\frac{1}{4}\mathcal{F}_{\mu\nu}^i(M^{-1})_{ij}\mathcal{F}^{\mu\nu j} \quad (3.43)$$

$$\mathcal{L}_4 = -\frac{1}{12}H_{\mu\nu\rho}H^{\mu\nu\rho} \quad (3.44)$$

Here we closely follow the notation of [49] and [64]. The next step is to perform a duality transformation, which replaces the field  $B_{\mu\nu}$  by a scalar field  $\chi$ . This is achieved by first forming the  $B_{\mu\nu}$  equation of motion

$$\partial_\mu(\sqrt{-g} e^{-2\phi} H^{\mu\nu\rho}) = 0 \quad (3.45)$$

and solving it by setting

$$\sqrt{-g} e^{-2\phi} H^{\mu\nu\rho} = \gamma \epsilon^{\mu\nu\rho\lambda} \partial_\lambda \chi \quad (3.46)$$

where  $\chi$  is the “axion” and  $\gamma$  is a constant to be fixed later. In the language of differential forms,

$$H = \gamma e^{2\phi} * d\chi \quad (3.47)$$

or, using  $H = dB - \frac{1}{2}\eta_{ij}A_\lambda^i \mathcal{F}^j$ ,

$$dB = \frac{1}{2}\eta_{ij}A_\lambda^i \mathcal{F}^j + \gamma e^{2\phi} * d\chi \quad (3.48)$$

The Bianchi identity ( $d^2 B = 0$ ) now turns into the  $\chi$  field equation

$$\frac{1}{2}\eta_{ij}\mathcal{F}_\lambda^i \mathcal{F}^j + \gamma d(e^{2\phi} * d\chi) = 0 \quad (3.49)$$

or, in terms of components, (choosing a convenient value for  $\gamma$ )

$$\partial_\mu(e^{2\phi}\sqrt{-g} g^{\mu\nu}\partial_\nu\chi) - \frac{1}{8}\eta_{ij}\epsilon^{\mu\nu\rho\lambda}\mathcal{F}_{\mu\nu}^i\mathcal{F}_{\rho\lambda}^j = 0, \quad (3.50)$$

This is an equation of motion if we replace the  $\mathcal{L}_4$  term in  $S^{(4)}$  by

$$S_\chi = - \int dx \sqrt{-g} \left( \frac{1}{2} e^{2\phi} g^{\mu\nu} \partial_\mu \chi \partial_\nu \chi + \frac{1}{4} \chi \mathcal{F} \cdot \tilde{\mathcal{F}} \right), \quad (3.51)$$

where

$$\mathcal{F} \cdot \tilde{\mathcal{F}} \equiv \frac{1}{2\sqrt{-g}} \epsilon^{\mu\nu\rho\lambda} \mathcal{F}_{\mu\nu}^i \eta_{ij} \mathcal{F}_{\rho\lambda}^j. \quad (3.52)$$

Let us briefly recapitulate the steps we have taken to modify the four dimensional action in the Einstein frame. The field strength  $H_{\mu\nu\lambda}$  appearing in  $\mathcal{L}_4$  is traded for the pseudoscalar axion,  $\chi$ . The resulting action (3.51) contains not only the kinetic energy term of the axion, but also the topological term which is like the  $\theta$  dependant term of the Yang-Mills action if the VEV of  $\chi$  is identified with that parameter.

Let us now regroup the terms in the dual action in the following way:

$$\tilde{S}^{(4)} = \int_M dx \sqrt{-g} (R + \mathcal{L}_2) + S_D + S_F, \quad (3.53)$$

where

$$S_D = -\frac{1}{2} \int_M dx \sqrt{-g} g^{\mu\nu} \left( \partial_\mu \phi \partial_\nu \phi + e^{2\phi} \partial_\mu \chi \partial_\nu \chi \right) \quad (3.54)$$

$$S_F = -\frac{1}{4} \int_M dx \sqrt{-g} \left( e^{-\phi} \mathcal{F}^2 + \chi \mathcal{F} \cdot \tilde{\mathcal{F}} \right) \quad (3.55)$$

and  $\mathcal{F}^2 \equiv g^{\mu\rho} g^{\nu\lambda} \mathcal{F}_{\mu\nu}^i (M^{-1})_{ij} \mathcal{F}_{\rho\lambda}^j$ . Note that  $\tilde{S}^{(4)}$  contains the usual Einstein-Hilbert action and the part coming from kinetic energy term of the M-matrix. We have rearranged the actions coming

from dilaton kinetic energy, gauge field part and the axionic part (together with the 'topological' term) to define  $S_D$  and  $S_F$  so that dilaton and axion are put together and the gauge field kinetic energy along with the topological term are clubbed together. This is very useful to study the S-duality properties of the action. In order to describe the  $SL(2, R)$  symmetry of the dilaton and axion kinetic terms, let us introduce a complex modular parameter (recall the case of Yang-Mills)

$$\tau = \chi + ie^{-\phi}, \quad (3.56)$$

which has the nice property that under a linear fractional transformation

$$\tau \rightarrow \frac{a\tau + b}{c\tau + d} \quad (3.57)$$

the combination

$$\frac{g^{\mu\nu} \partial_\mu \tau \partial_\nu \bar{\tau}}{(\text{Im } \tau)^2} = g^{\mu\nu} (\partial_\mu \phi \partial_\nu \phi + e^{2\phi} \partial_\mu \chi \partial_\nu \chi) \quad (3.58)$$

is invariant. It immediately follows that

$$S_D = -\frac{1}{2} \int_M dx \sqrt{-g} \frac{g^{\mu\nu} \partial_\mu \tau \partial_\nu \bar{\tau}}{(\text{Im } \tau)^2}. \quad (3.59)$$

Now we consider the gauge field action,  $S_F$ . Notice that the  $SL(2, R)$  transformations give rise to an electric-magnetic duality rotation. Let us define

$$\mathcal{F}_{\mu\nu}^\pm = M\eta \mathcal{F}_{\mu\nu} \pm i\tilde{\mathcal{F}}_{\mu\nu}. \quad (3.60)$$

Then, using the identity  $\mathcal{F}^{+\mu\nu} M^{-1} \mathcal{F}_{\mu\nu}^- = 0$ , we can express  $S_F$  as

$$S_F = -\frac{1}{16i} \int_M dx \sqrt{-g} \left( \tau \mathcal{F}^{+\mu\nu} M^{-1} \mathcal{F}_{\mu\nu}^+ - \bar{\tau} \mathcal{F}^{-\mu\nu} M^{-1} \mathcal{F}_{\mu\nu}^- \right). \quad (3.61)$$

The  $\mathcal{A}_\mu$  equation of motion is

$$\nabla^\mu (\tau \mathcal{F}_{\mu\nu}^+ - \bar{\tau} \mathcal{F}_{\mu\nu}^-) = 0 \quad (3.62)$$

and the Bianchi identity is

$$\nabla^\mu (\mathcal{F}_{\mu\nu}^+ - \mathcal{F}_{\mu\nu}^-) = 0. \quad (3.63)$$

To exhibit  $SL(2, R)$  symmetry it is necessary to have  $\mathcal{A}_\mu$  transform at the same time as  $\tau$ . The appropriate choice is to require that  $\mathcal{F}_{\mu\nu}^\pm$  transform as modular forms as follows

$$\mathcal{F}_{\mu\nu}^+ \rightarrow (c\tau + d) \mathcal{F}_{\mu\nu}^+, \quad \mathcal{F}_{\mu\nu}^- \rightarrow (c\bar{\tau} + d) \mathcal{F}_{\mu\nu}^-. \quad (3.64)$$

This implies that

$$\tau \mathcal{F}_{\mu\nu}^+ \rightarrow (a\tau + b) \mathcal{F}_{\mu\nu}^+, \quad \bar{\tau} \mathcal{F}_{\mu\nu}^- \rightarrow (a\bar{\tau} + b) \mathcal{F}_{\mu\nu}^-. \quad (3.65)$$

Thus the equation of motion (3.62) and the Bianchi identity (3.63) transform into linear combinations of one another and are preserved. In particular, the negative of the unit matrix sends  $\mathcal{F}_{\mu\nu}^\pm \rightarrow -\mathcal{F}_{\mu\nu}^\pm$ . This result is acceptable if we identify the symmetry as  $SL(2, R)$ . Note that  $SL(2, R)$  is not a symmetry of the action. The transformation in (3.64) is a nonlocal transformation of  $\mathcal{A}_\mu$ , and such transformations can do strange things to the action. For example, the total derivative  $\mathcal{F} \cdot \tilde{\mathcal{F}}$  transforms into an expression that is not a total derivative.

Thus far we have focused the attention to dilaton-axion system and the gauge field part of the action. The explicit checks show that the rest of the equations of motion are invariant under S-duality transformation. While checking the invariance of the Einstein equation we must ensure that the contribution of  $S_F$  to the energy-momentum tensor is  $SL(2, R)$  invariant. After a short calculation one finds that only terms of the structure  $e^{-\phi} \mathcal{F}^+ \mathcal{F}^-$  survive, and these are invariant since  $e^{-\phi} \rightarrow |c\tau + d|^{-2} e^{-\phi}$ . The symmetry of the equations of motion is  $SL(2, R)$ . Notice that the axion couples to the topological density term, product of  $\mathcal{F}$  and its dual. We can argue

qualitatively that the part of the  $SL(2, R)$  group which gives rise to the translation symmetry of the axion (VEV of  $\chi$  is the  $\theta$  angle) should break down to discrete group of translations due to the instanton effects. A more careful analysis is necessary [42] to show that  $SL(2, R)$  breaks to  $SL(2, Z)$ .

The low energy string effective action, in four dimensions, contains graviton, antisymmetric tensor, dilaton and nonabelian gauge bosons. Furthermore, the Poincare dual of the three form field strength is a pseudoscalar and this field can be identified as the axion. One can combine dilaton and axion to form a doublet of the S-duality group  $SL(2, R)$ . It was argued [61, 62] that S-duality is an exact symmetry of the string theory. Schwarz and Sen [63] provided a general formulation of S-duality in string theory. Indeed the heterotic string compactified on  $T^6$  has the effective action of  $N = 4$  supersymmetric theory. How one can test S-duality in this case. One of the important results in this direction was first derived by Sen [96] when he showed that there are certain dyonic states in the theory whose existence can be demonstrated using S-duality transformations on heterotic string actions. These states precisely coincide with the ones we expect from Montonen-Olive conjecture. The theory has electrically charged states and magnetically charged states and each is 28-dimensional vector for the heterotic string. Due to nonrenormalization theorem of  $N = 4$  supersymmetric theory, the electric charges are not renormalised. Moreover, the spectrum of the magnetic charges are fixed by the generalized Dirac quantization condition; the magnetic charges are not renormalised either. Thus, spectrum of these charges will be same as in the tree level theory. Indeed, the multimoduli could be computed for the heterotic string [96]. In fact, the study of nonperturbative aspects of supersymmetric Yang-Mills theories took new directions through the works of Seiberg and Witten [66] in sequel to Sen's work.

It is interesting to look for extended objects which appear as solution to equations of motion of string effective action. Simplest extended object is a string which is one dimensional. Let us denote the worldsheet coordinates of this string as  $\xi^0$  and  $\xi^1$  and the spacetime coordinates as  $\{x^\mu\}$ . This should appear as solution to string effective action. Suppose, we consider a frame where  $(\xi^0, \xi^1)$  lie along the spacetime coordinates  $(x^0, x^1)$  respectively. We look for a 'spherically symmetric' solution such that the solution is static and it depends only on the magnitude of the transverse distance,  $r = \sqrt{y_1^2 + \dots + y_8^2}$  where  $x^2 \dots x^9$  are denoted as  $y_i$ 's. The effective action has graviton, dilaton and antisymmetric tensor fields. In the Einstein frame the action has the form

$$S_E = \frac{1}{\kappa^2} \int d^{10}x \sqrt{-g} [R - \frac{1}{2}(\partial\phi)^2 - \frac{1}{12}e^{-\phi}H^2] \quad (3.66)$$

The macroscopic string solution which was identified with the heterotic string [67] is obtained for following background configurations

$$ds^2 = f^{-\frac{3}{4}}(-dt^2 + (dx^1)^2) + f^{\frac{1}{4}}dy^i dy^i \quad (3.67)$$

$$B_{01} = \frac{1}{f} \quad (3.68)$$

The rest of the components of  $B_{\mu\nu}$  are set to zero and

$$f = 1 + \frac{q}{3r^6} \quad (3.69)$$

Here  $Q$  is the charge carried by the string and it is associated with antisymmetric tensor field. The field equations one needs to satisfy are: Einstein equation, dilaton field equation and axionic charge conservation which follow from field equation of  $H$ . If we look at field equation carefully there is a delta-function singularity at  $r = 0$  in the Laplace equation  $\nabla^2 f$ . Therefore, it was proposed [67] to resolve this problem by introducing a source for the string which will be the  $\sigma$ -model action

$$S_\sigma = \frac{-T}{2} \int d^2\xi [\partial^a X^\mu \partial_a X^\nu G_{\mu\nu} + \epsilon^{ab} \partial_a X^\mu \partial_b X^\nu B_{\mu\nu}] \quad (3.70)$$

Here of course the metric  $G_{\mu\nu}$  is the string frame metric. This is the string solution carrying 'electric' charge and this charge can be obtained from the conservation law. Indeed,  $q = \kappa^2 T / \omega_7$

where  $\omega_7$  refers to the volume of  $S^7$ . In the supersymmetric case, there are BPS saturating solutions and here mass per unit length is equal to the charge.

In four dimensions the dual of electromagnetic field tensor is also a two form, thus if we have point particles, the dual objects are point-like ('t Hooft-Polyakov monopoles look point like at large distances). However, if we have a string in ten dimensions it couples to 3-form field strength the dual of that field strength is 7-form. Therefore, the solitonic object for the string is a 5-brane, extended in five spatial dimensions. In fact the p-brane solutions were found in sequel to the string solutions [68]. As in case of monopole solution, we do not have magnetic source term while looking for field equations (W-bosons carry electric charge), the solitonic five-branes solutions are derived without adding a source term. Moreover, if  $e_2$  is 'electric' charge of the string and  $g_6$  is 'magnetic' charge of soliton, the Dirac quantization condition is

$$e_2 g_6 = 2\pi n \quad (3.71)$$

One has to be careful in deriving strong-weak duality relation here. The coupling constant is determined in terms of dilaton expectation value. The relations are  $e_2 = e^{\phi_0/2}$  and  $g_6 = e^{-\phi_0/6}$ . There are special type of extended objects, the Dp-brane (D-branes), which carry R-R charges [69]. The type II theories admit gauge fields from the RR sector. The corresponding effective contain these fields. If one look for p-brane solutions with these gauge fields: strings, membranes and so on, they have interesting properties. These are hypersurfaces or spacetime defects on which the open strings can end. In D-dimensions, if there is a Dp-brane, there are Neumann boundary conditions satisfied in  $(p+1)$ -directions, these are the directions of the worldvolume coordinates of Dp-brane and we have Dirichlet boundary conditions along the remaining transverse directions that is  $(D - p - 1)$  coordinates. Written explicitly,

$$\partial_\sigma X^\mu = 0, \text{ for } \mu = 0, \dots, p \quad (3.72)$$

$$X^\mu(\sigma = 0, \pi) = a_0^\mu, \text{ for } \mu = p+1, \dots, 9 \quad (3.73)$$

A Dp-brane will couple to  $(p+2)$ -form RR field strength; therefore, D0-brane is a particle, D1-brane is a string and so on. The corresponding fermions satisfy boundary conditions in accordance with the bosonic fields in order to maintain the worldsheet supersymmetry. The BPS saturated solutions, then preserve half of the supersymmetry. From our earlier discussions, we note that type IIA admits D0-brane and D2-brane (their dual objects too) and IIB string has D-string, D3-brane and D-instantons, along with the duals. Thus, we conclude that IIA has even D-branes and odd D-branes belong to IIB theory. Of course, we are discussing the 10 dimensional case. The D-branes are dynamical objects and there are excitation of such extended objects since open string ends are attached to them.

Consider a situation when two D-branes are separated from each other. Since open string ends can get attached to this surface, they will be connected by open string/strings. The farther apart the two branes, it will cost more energy to stretch the open string. More interesting is the configuration when D-branes lie on top of each other. Then we can visualise an open string starting from a brane and ending on it, open string starting from one brane and ending on another coincident brane. In this situation we have massless states since there is no stretching of strings. Open strings contain massless vector state in their spectrum. One can incorporate nonabelian gauge symmetry for such a theory by introducing the Chan-Paton factors. We can imagine a scenario where a quark belonging to representation  $i$  of  $U(n)$  is attached to one end of the string and an antiquark in representation  $\bar{j}$  attached to the other end. Thus the gauge field will carry index  $i$  and  $j$  like usual Yang-Mills fields and these are called Chan-Paton factors. This characteristic of open string turned out to be useful when we consider coincident D-branes. Therefore, if there are  $N$  coincident branes, we get  $U(N)$  Yang-Mills action, in fact we get supersymmetric gauge theory on the worldvolume of the brane.

Let us discuss some of the implications of dualities in the context of the branes we just introduced. The experience from monopole solution is that the charged particle couples to the field strength tensor and the soliton couples to the dual tensor in four dimensions. In ten dimensions, the solitonic counter part of string is five brane and we saw that couplings are not really reciprocals of each

other. If we consider six spacetime dimensions, then we note that dual of 3-form field strength is also another 3-form tensor and string couples to this tensor. Therefore, the conjecture is that in six dimensions there is string/string duality. If there is a fundamental string the solitonic counter part is a string too and their coupling constants satisfy the reciprocal relation. For simplicity, consider a six dimensional reduced action, with only metric, antisymmetric tensor field and the dilaton [70].

$$I_6 = \frac{1}{2\kappa^2} \int d^6x \sqrt{-G} e^{-\phi} [R_G + (\partial\phi)^2 - \frac{1}{12} H^2] \quad (3.74)$$

Where  $G_{MN}$  is six dimensional metric in string frame and  $H_{NMP}$  is the 3-form field strength associated with  $B_{MN}$  and it is understood that  $H$  is defined up to Chern-Simons terms. We can go over to Einstein metric by the relation  $G_{MN} = e^{\phi/2} g_{MN}$ ;  $\phi$  being the dilaton in six dimensions. Let us consider the dual six dimensional action

$$\tilde{I}_6 = \frac{1}{2\kappa^2} \int d^6x \sqrt{-\tilde{G}} e^{-\tilde{\phi}} [\tilde{R}_{\tilde{G}} + (\partial\tilde{\phi})^2 - \frac{1}{12} \tilde{H}^2] \quad (3.75)$$

Here  $\tilde{\phi}$  is the corresponding dilaton and  $\tilde{H}$  is the field strength of the  $\tilde{B}$ , 2-form potential of the dual theory. The two actions (3.74) and (3.75) are related if we identify

$$\phi = -\tilde{\phi} \quad \text{and} \quad \tilde{H} = e^{-\phi} * H \quad (3.76)$$

The two metric being identified to be equal. Here  $*$  stands for Hodge dual. Note that just as in case of gauge field kinetic energy term in four dimensions is conformally invariant, the  $H^2$  term is also conformally invariant in six dimensions and it is immaterial which metric we use while taking Hodge dual. As noted earlier, the fundamental string solution with action (3.74) can be obtained by adding a  $\sigma$ -model source term with coupling of the G and B backgrounds. The solution is given by

$$ds^2 = (1 - \frac{q^2}{r^2}) [-dt^2 + (x^1)^2 + (1 - \frac{q^2}{r^2})^{-2} dr^2 + r^2 d\Omega_3^2] \quad (3.77)$$

$$e^{\phi} = 1 - \frac{q^2}{r^2} \quad (3.78)$$

$$e^{-\phi} * H_3 = 2q^2 \epsilon_3 \quad (3.79)$$

with

$$q^2 = \frac{\kappa^2 T}{\Omega_3} \quad (3.80)$$

Of course we have the BPS saturated mass relation

$$M = T < e^{\frac{\phi}{2}} > \quad (3.81)$$

Therefore, the mass density gets heavier as string coupling proceeds towards strong coupling domain. The source free action (3.74) also admits solitonic string which is nonsingular and the solution is

$$ds^2 = -dt^2 + (dx^1)^2 + (1 - \frac{\tilde{q}^2}{r^2})^{-2} dr^2 + r^2 d\Omega_3^2 \quad (3.82)$$

$$e^{-\phi} = 1 - \frac{\tilde{q}^2}{r^2} \quad (3.83)$$

$$H_3 = 2\tilde{q}^2 \epsilon_3 \quad (3.84)$$

Where  $\tilde{q}^2 = \frac{q^2 \tilde{T}}{\Omega_3}$  The mass density is

$$\tilde{M} = \tilde{T} < e^{-\frac{\phi}{2}} > \quad (3.85)$$

In the weak coupling regime this string is heavier as one expects of a solitonic string. Notice that the solitonic string differs from the fundamental string by the replacement  $\phi \rightarrow -\phi, G_{MN} \rightarrow$

$\tilde{G}_{MN}, H \rightarrow \tilde{H} = e^{-\phi} * H, \alpha' \rightarrow \tilde{\alpha}'$ . The Noether charge and the topological 'magnetic' charge are respectively given by

$$e_2 = \frac{1}{\sqrt{2}\kappa} \int_{S^3} *H_3, \text{ and } g_2 = \frac{1}{\sqrt{2}\kappa} \int_{S^3} H_3 \quad (3.86)$$

The Dirac quantization rule for charges:  $e_2 g_2 = 2\pi n$  gets translated to relation between tensions. Moreover, the fundamental string and dual string saturate Bogomolnyi bound for mass densities and break half of the supersymmetry as expected. These solutions have the interpretation of being limiting cases of more general solutions. They can be viewed as extreme mass equals charge limit of two-parameter black string solutions.

Again the question arises where can we test the string/string duality? It has been conjectured that [71, 72, 73] heterotic string compactified on  $T^4$  is S-dual to type IIA theory compactified on  $K_3$ . When heterotic string is compactified on  $T^4$ , the theory has charged states saturating Bogomolnyi bound. On the IIA side, elementary string states are neutral since the gauge fields arise from RR sector. Moreover, for type IIA, the analysis of the Bogomolnyi formula tells us the charged states (under gauge fields) have their masses as  $\frac{1}{g_{str}}$ , implying that these are solitonic states. The duality between heterotic and type IIA is understood in the following sense [74, 75]: In type IIA theory, there are nonsingular soliton solutions and these carry quantum numbers of fundamental heterotic string. The properties of those strings are consistent with those of the heterotic string. On the other hand the heterotic string admits solitonic solutions carrying the quantum numbers of type IIA string. Moreover, we know that the moduli of heterotic string compactified on  $T^4$  parametrize the coset  $\frac{O(4,20)}{O(4) \times O(20)}$ . When type IIA is compactified on  $K_3$ , the moduli also turns out to be exactly the same. Therefore, there is a very good evidence for this heterotic - type IIA duality conjecture. Another duality relation, that has been verified, is toroidal compactification of IIA and IIB theory via T-duality. Again the simplest one being compactification on  $S^1$ . If one theory is compactified on circle of radius  $R$ , it is equivalent to the other theory compactified on circle of reciprocal radius [76], although in ten dimensions these are two different theories. Some of the important consequences of S-duality can be examined in type IIB theory. It is conjectured that type IIB theory is self-dual and the effective action can be cast in a manifestly  $SL(2, \mathbb{Z})$  invariant form. We shall study this aspect in the next section. The two heterotic strings i.e.  $SO(32)$  and  $E_8 \times E_8$  when compactified on  $S^1$  are T-dual to each other in the reciprocal radius sense that one theory compactified on a circle of radius  $R$  is equivalent to the other which is compactified on a circle of radius  $\frac{1}{R}$ . Finally, we comment that heterotic string with  $SO(32)$  gauge group is S-dual to type I theory with  $SO(32)$  group. The heterotic string effective action, with  $SO(32)$  gauge group has the following form

$$S_{het} = \int d^{10}x \sqrt{-g} [R - \frac{1}{8}(\partial\phi)^2 - \frac{1}{4}e^{-\frac{\phi}{2}} \text{Tr}(F_{\mu\nu})^2 - \frac{1}{12}e^{-\frac{\phi}{2}} H^2] \quad (3.87)$$

Here  $F_{\mu\nu}$  is the nonabelian field strength and  $H = dB$ . We work in the Einstein frame as it is the most convenient frame to study S-duality properties, since this metric remains invariant under S-duality. This action is obtained after rescaling the backgrounds and the slope parameters. The type I string has graviton and dilaton coming from the closed string NS sectors and closed string RR sector gives the antisymmetric tensor. The gauge fields come from NS sector of the open string and they have to be in the adjoint representation of  $SO(32)$ . Again with appropriate scalings the effective action can be brought to the following form

$$S_I = \int d^{10}x \sqrt{-\bar{g}} [\bar{R} - \frac{1}{8}(\partial\bar{\phi})^2 - \frac{1}{4}e^{\frac{\bar{\phi}}{4}} \text{Tr}(\bar{F}_{\mu\nu})^2 - \frac{1}{12}e^{\frac{\bar{\phi}}{2}} \bar{H}^2] \quad (3.88)$$

Here all the fields of type I theory are defined with 'bar' to distinguish from those of heterotic string theory and the metric is in Einstein frame. Now, the comparison between the two actions shows that they will be identical if

$$\phi = -\bar{\phi}, \quad g_{\mu\nu} = \bar{g}_{\mu\nu}, \quad H_{\mu\nu\rho} = \bar{H}_{\mu\nu\rho}, \quad A_\mu = \bar{A}_\mu \quad (3.89)$$

Thus, if we compare the two actions, (3.87) and (3.88), we see that the two theories are related to each other by strong-weak duality in 10-dimensions, since  $g_{str}^2 = e^\phi$ . There are host of duality relations among various string theories in diverse dimensions; we refer the interested reader to large number of review articles in this area.

## 4 M-theory and Unified String Dynamics

We have briefly introduced some of the essential features of string theory and their symmetry properties. There are five perturbatively consistent string theories and one of their most attractive attributes is that they describe quantum gravity which is perturbatively finite and unitary. The dualities are powerful symmetry properties which provide important information about intimate connections between string theories. We have seen that one string theory, in a spacetime dimension, is related to another string theory either through T-duality or by the S-duality. When two theories are S-dual to each other, we can study strong coupling regime of one theory by going over to the weak, perturbative domain of its dual theory. Therefore, the nonperturbative aspects of some of the string theories could be investigated by these powerful tools. However, we still have five string theories. Therefore, the natural goal is to search for a theory which will provide a unified description of all the five string theories. The zero slope limits of the string theories yield all the known 10-dimensional supergravity theories. However, there is the  $D = 11$  supergravity theory consisting of graviton and 3-form potential, endowed with total 128 bosonic degrees of freedom, and the 128 fermionic degrees of freedom. It was shown several years ago [77] that compactification of 11-dimensional theory on a circle gives rise to  $N = 2$  supergravity theory in 10-dimensions. It was not possible to establish any relation between the 11-dimensional theory and any string theory for a long time. The connection of  $N = 2$ , 10-dimensional supergravity with string theory is rather transparent since the supergravity actions can be obtained in the zero slope limit of corresponding type II string theories. There was no string theory that could be related in some such limit to 11-dimensional supergravity. Therefore, if 11-dimensional supergravity were to have any connection with one of the string theories, then only the nonperturbative regime of a theory will show the inter-relation. Moreover, when one views from the 11-dimensional perspective, the supergravity theory does not have any small parameter, like  $e^\phi$ , in string theory, which can be chosen to take small value as an expansion parameter.

The connection between type IIA string theory and 11-dimensional supergravity were recognised by Witten [73] and Townsend [78] following the developments in string dualities. The massless bosonic sector of the type IIA theory, we might recall from our discussions of Section II, consists of diaton,  $\phi$ , graviton,  $G_{\mu\nu}$  and gauge field,  $A_\mu$ , antisymmetric tensor,  $C_{\mu\nu\lambda}$  coming from the NS and Ramond sectors respectively. The effective action of type IIA theory

$$S_{IIA} = \frac{1}{2\kappa_{10}^2} \int d^{10}x \sqrt{-G} [e^{-\phi} (R + (\partial\phi)^2 - \frac{1}{12} H^2) - (\frac{1}{4} F^2 - \frac{1}{48} F_4'^2)] - \frac{1}{4\kappa_{10}^2} \int F_4 \wedge F_4 \wedge A \quad (4.1)$$

We have suppressed the Lorentz indices of the field strengths and we shall define them now:  $R$  is the scalar curvature,  $H_{\mu\nu\rho}$  is the field strength of  $B_{\mu\nu}$  from the NS sector,  $F_{\mu\nu}$  is the field strength of RR gauge potential  $A_\mu$  and in form notations, 4-form field strength,  $F_4' = dC_3 + A \wedge dB$ ;  $C_3$  being the 3-form potential coming from the RR sector and  $B$  is the 2-form potential whose field strength is  $H$ . Last term in (4.1) is the Chern-Simons term. and  $F_4 = dC_3$  is the antisymmetric 4-form field strength of potential  $C_3$ . A few remarks are in order at this point: the metric used in action (4.1) is the string frame metric. Note that the factor  $e^\phi$  multiplies only  $R$  and  $H^2$  piece; fields coming from the NS sector. The reason is that in the worldsheet supersymmetric formulation of NSR type II theories the R-R sector fields through local worldsheet interactions (in NS sector the worldsheet fields couple to potentials), couple via bilinears of spin fields (in fact to field strengths). As a consequence, there are cuts and the usual arguments that tree level term starts with  $\frac{1}{g_{str}}$  does not go through. Thus we see this mismatch of  $e^\phi$  between NS and RR fields in the effective action.



Now, it is easy to see that this theory will admit D0-brane and D2-brane and their duals will be D6-brane and D4-brane from RR sector and a string and its dual five brane from the NS sector. Let us consider the bosonic part of the eleven dimensional supergravity action

$$S_{11} = \frac{1}{2\kappa_{11}^2} \int d^{11}x \sqrt{-\tilde{G}} [\tilde{R} - \frac{1}{48} \tilde{F}_4^2] - \frac{1}{12\kappa_{11}^2} \int \tilde{C}_3 \wedge \tilde{F}_4 \wedge \tilde{F}_4 \quad (4.2)$$

Here the field with tilde belong to bosonic components of 11-dimensional supergravity. Let us compactify one of the spatial dimensions on  $S^1$ , following the procedure outlined in the last section. There will be a gauge field and a scalar field, when the metric is expressed is decomposed in terms of the metric of the 10-dimensional theory. The 3-form potential will decompose into a 3-form potential but with additional piece according to the procedure of [48, 49] and a two form potential will appear as well. It is most convenient to express the 11-dimensional metric in the following form

$$\tilde{G}_{MN} = e^{-\frac{1}{3}\phi} \begin{pmatrix} G_{\mu\nu} + e^\phi A_\mu A_\nu & e^\phi A_\mu \\ e^\phi A_\nu & e^\phi \end{pmatrix} \quad (4.3)$$

The dimensional reduction of (4.2) goes over exactly to the type IIA action (4.1). Note that if we had not adopted this form of the decomposition of the 11-dimensional metric with the over all factor of  $e^{-\frac{1}{3}\phi}$  and the factors of  $e^\phi$  in various places inside the matrix; but had compactified on a circle of radius, say,  $\mathcal{R}$ ; we would have obtained a reduced action with 10-dimensional metric, the moduli  $\mathcal{R}$  and the antisymmetric tensor potentials (2-form and 3-form) with appropriately modified C-S terms. The resulting action in ten dimensions would need some field redefinitions to match with the type IIA action. Let us see how the radius of compactification  $R_{11}$  is related to type IIA string coupling constant  $g_s^{(A)}$ . Note from (4.3) that  $R_{11}^2 = (e^{\frac{2}{3}\phi})^2$  and by definition  $e^\phi = (g_{str}^{(A)})^2$ . Therefore, we conclude that

$$R_{11} = (g_{str}^{(A)})^{\frac{2}{3}} \quad (4.4)$$

Therefore, in the perturbative regime of the type IIA theory, the radius of compactification of the 11-dimensional theory is very small. When we want to go over to the decompactification regime i.e. large radius limit of 11-dimensional theory, we can't realise that domain since it is the strong coupling phase of the type IIA theory and perturbation theory does not provide any clue for the existence of the 11th dimension in the ten dimensional theory. The correspondence established between type IIA theory and 11-dimensional theory is at the level of the effective action. The 11-dimensional supergravity has a 3-form potential in the bosonic sector and the natural extended object is a membrane. The 10-dimensional theory admits a string as a fundamental object and supergravity action is zero slope limit of the string theory. How can one establish the relation between membrane and the string? The idea of double dimensional reduction provides an important clue. One can envisage a situation where we start from a membrane in eleven dimensions and compactify 11th dimension on a circle. Then, according to the prescription of double dimensional reduction [79], the membrane wraps around the compact direction so that the end result is the ten dimensional string.

We have described in the previous section how one can establish connections among the five string theories various dimensions through duality transformations in different spacetime dimensions; although there are five distinct ten dimensional theories when viewed in the perturbative frame work. The 11-dimensional theory is also recognised to play an important role in string dynamics. It is believed that there is an underlying fundamental theory, yet to be discovered, so that the manifestations of the theory in its various phases are realized through the string theories. It is postulated that in the low energy limit, we should derive the 11-dimensional supergravity action as an effective theory. The unknown fundamental theory is named U-theory. Since the 11-dimensional theory naturally admits membrane as a fundamental extended solution, it has been argued that the underlying fundamental theory is a theory of membranes. The M-theory is taken to be the underlying theory. We shall illustrate, with a few examples, that starting with an eleven dimensional theory with membrane, how one can obtain a host of relations about the structure of branes in

various string theories.

Since the BPS states do not get any quantum corrections, it is interesting to look for BPS states and then propose tests for the theory. When we compactify M-theory on a circle, the momenta in that direction will be quantized and we shall get towers of KK massive states. These states will fall into representations of the 11-dimensional supergravity. In fact they are BPS states. In the KK reduction, the charge of a state (in the lower dimensions) is related to the momentum along the compact direction (thus automatically quantized) and in some suitable units the charge is proportional to  $\frac{m}{R_{11}}$ ,  $m$  being an integer. This is the charge associated with the gauge field  $A_\mu$  as a result of  $S^1$  compactification (4.3). From the type IIA point of view this charge is that of gauge field coming from RR sector and the whole tower should exist as BPS state. We know already that elementary string states are RR-charge neutral and those massive towers belong to RR sector. We can identify the state with unit charge,  $m = 1$  as a D0-brane of type IIA theory. The open string ends can get attached to D0-brane and act as the collective coordinates to give excitations. One can show that IIA theory has those BPS states belonging to the ultra-short multiplets and these also correspond to the states counting done from M-theory side. Therefore, we notice that duality between type IIA theory and M-theory is established for such states. In case of  $m > 1$ , the test is not so simple. One of the properties of BPS states is that the binding energy for composite BPS states is zero. That means, if we have a single D0-brane, a BPS state with  $m$  units of charge, we can't distinguish it from collection of  $m$  BPS particles each carrying unit charge. Thus a test for the general case is rather difficult.

The relation between M-theory and type II theories can be established by exploiting the duality relations. Note that type IIA and type IIB theories are T-dual to each other when one of the directions is compactified. Since M-theory with one compact direction,  $S^1$ , is related to type IIA, therefore, M-theory with two compact dimensions, compactified on  $T^2$ , is expected to be intimately connected [80] to type IIB with one direction compactified to  $S^1$ . We shall see that one needs to exploit the  $SL(2, Z)$  S-duality symmetry of type IIB theory in this context [81]. The type IIB theory has graviton, 2-form antisymmetric potential,  $B_{\mu\nu}^{(1)}$  and dilaton,  $\phi$  in the NS sector and 2-form potential,  $B_{\mu\nu}^{(2)}$ , axion,  $\chi$  and 4-form potential,  $D_{\mu\nu\rho\lambda}$  in the RR sector; the field strength of D-field is self dual. For our purpose, it suffices to drop the D-field from considerations presently. The action is

$$S_{str} = \frac{1}{\kappa^2} \int d^{10}x \sqrt{-G} \left[ e^{-\phi} (R + (\partial\phi)^2 - \frac{1}{12} H^{(1)2}) - \frac{1}{2} (\partial\chi)^2 - \frac{1}{12} \chi^2 H^{(1)2} - \frac{1}{6} \chi H^{(1)} \cdot H^{(2)} - \frac{1}{12} H^{(2)2} \right] \quad (4.5)$$

This action is written in the string frame metric. It is useful to go over to the Einstein frame by the conformal transformation. Furthermore, to write the Einstein frame action in a manifestly  $SL(2, Z)$  invariant form, let us define

$$\mathcal{M} = \begin{pmatrix} \chi^2 + e^{-\phi} & \chi e^\phi \\ \chi e^\phi & e^\phi \end{pmatrix}, \quad \mathbf{H} = \begin{pmatrix} H^{(1)} \\ H^{(2)} \end{pmatrix} \quad (4.6)$$

Then the action,

$$S_E = \frac{1}{2\kappa^2} \int d^{10}x [R_g + \frac{1}{4} \text{Tr}(\partial_\mu \mathcal{M} \partial^\mu \mathcal{M}) - \frac{1}{12} \mathbf{H}^T \mathcal{M} \mathbf{H}] \quad (4.7)$$

This action is invariant under the transformations

$$\mathcal{M} \rightarrow \Lambda \mathcal{M} \Lambda^T, \quad \mathbf{H} \rightarrow (\Lambda^T)^{-1} \mathbf{H} \quad \text{and} \quad g_{\mu\nu} \rightarrow g_{\mu\nu} \quad (4.8)$$

If one looks for a string solution in this theory then the solutions will be of three kinds: strings carrying NS charge, strings with RR charge and ones with both NS and RR charge. The procedure adopted in [81] is as follows: first look for a string solution with NS charge such that asymptotic values of axion,  $\chi_0 = 0$  and that of dilaton  $\phi_0 = 0$ . In the language of complex moduli introduced earlier, asymptotic value of  $\tau_0 = i$ . Moreover, one starts with  $H^{(2)} = 0$ , since one is looking for a

string carrying NS charge only. Next introduce a specific  $SL(2, Z)$  transformation such that the resulting string carries both types of charges; the relevant matrix is

$$\Lambda = \frac{1}{\sqrt{q_1^2 + q_2^2}} \begin{pmatrix} q_1 & -q_2 \\ q_2 & q_1 \end{pmatrix} \quad (4.9)$$

Although the string carries both types, NS and RR charges, still the modulus preserves the asymptotic value,  $\tau_0 = i$ . Finally, introduce a general  $SL(2, Z)$  transformation so that  $\tau_0$  will take arbitrary value as a result of the duality transformation. The matrix is  $\Lambda = \begin{pmatrix} e^{-\phi_0/2} & \chi_0 e^{\phi_0/2} \\ 0 & e^{\phi_0/2} \end{pmatrix}$ . As a consequence of the  $SL(2, Z)$  transformation, not only we have strings which carry charges  $(q_1, q_2)$ , but also the tensions of the strings depend on these charges; after all these are BPS strings. The formula for the tension of string with  $(q_1, q_2)$  charges is

$$T_q = [e^{\phi_0}(q_2\chi_0 - q_1)^2 + e^{-\phi_0}q_2^2]^{\frac{1}{2}}T \quad (4.10)$$

where  $T$  is the tension of the NS string one started with for which  $\phi_0 = \chi_0 = 0$  i.e.  $\tau_0 = i$ . Since we consider  $SL(2, Z)$  transformations,  $q_1$  and  $q_2$  should be integers. For stable strings,  $(q_1, q_2)$  should be relatively prime; otherwise these string will decay into multiple strings. If we compactify this theory on  $S^1$ , more interesting results follow. The spectrum of the nine dimensional theory is governed by the mass formula

$$M_B^2 = \left(\frac{m}{R}\right)^2 + (2\pi R T_q n)^2 + 4\pi(N_L + N_R) \quad (4.11)$$

Here we have explicitly kept the tension term showing that when the string winds 'n' times it stretches by its perimeter and energy is obtained by multiplying the tension  $T_q$ . The last term is sum of contributions from left and right oscillators. The level matching condition tells us  $N_R - N_L = mn$ . The BPS saturating multiplets have either  $N_L = 0$  or  $N_R = 0$ ; ultrashort corresponds to both being zero. If we choose  $N_L = 0$ , then mass formula is (also level matching relation is used)

$$M_B^2 = (2\pi n R T_q + \frac{m}{R})^2 \quad (4.12)$$

We have a rich spectrum and these masses should remain protected from any quantum corrections. We can describe the same phenomena by compactifying the M-theory on  $T^2$ . There is membrane in M-theory with tension  $T_{11}$  and if it wraps  $m$  times on a torus of area  $A_{11}$ , then the contribution to mass will be of the form  $m A_{11} T_{11}$ . But the area is  $A_{11} = (2\pi R_{11})^2 \rho_2$ ,  $\rho_2$  being the modular parameter of the torus and  $\rho = \rho_1 + i\rho_2$  the area is computed using 11-dimensional metric. Since we are considering compactification of the M-theory on  $T^2$ , the wave function of the two dimensional Laplacian (corresponding to two coordinates on the torus) must satisfy periodicity property appropriate to the torus and the mass formula should be suitably generalised with respect to mass formula for a string compactified on a circle.

$$M_{11}^2 = [m(2\pi R_{11})^2 \rho_2 T_{11}]^2 + \frac{1}{R_{11}^2} [l_2^2 + \frac{1}{\rho_2^2} (l_1 - l_2 \rho_1)^2] \quad (4.13)$$

$l_1, l_2$  are integers which enter the mass formula as the contribution to the KK part since the two dimensional Laplacian  $-\partial_x^2 - \partial_y^2$  acting on the wave function

$$\psi_{l_1, l_2}(x, y) = \exp\left\{\frac{i}{R_{11}}[x l_2 + \frac{1}{\rho_2} y (l_1 - l_2 \rho_1)]\right\} \quad (4.14)$$

It is easy to see the periodicity property of the wave function by defining  $z = (x + iy)/2\pi R_{11}$ , since the invariance now translates to  $z \rightarrow z + 1$  and  $z \rightarrow z + \rho$ . In order to compare the above mass formula, obtained from M-theory with  $T^2$  compactification, with the corresponding one (4.12) from type IIB in 9-dimensions, we should recognize that (4.14) is derived using 11-dimensional metric. Therefore they could differ from each other by a multiplicative constant:  $M_{11} = C M_B$ . Now the

exact matching of the mass formula implies that the modular parameters of  $T^2$ , denoted as  $\rho$ , should be identical to the parameters of  $SL(2, Z)$ ,  $\tau$ . Thus the modular group appearing in  $T^2$  compactification of M-theory,  $SL(2, Z)$  is identical to the duality group,  $SL(2, Z)$ , of type IIB theory. The following relations should be satisfied for matching of (4.12) and (4.14)

$$R^{-2} = TT_{11}A_{11}^{\frac{3}{2}}, \text{ and } C^2 = \frac{2\pi R_{11}e^{-\phi_0/2}T_{11}}{T} \quad (4.15)$$

Since type IIA theory, in 9-dimensions, is related to IIB theory by T-duality, we can also get some insight into D0-branes in IIA theory. The mass spectrum of these (point) particles can be viewed from two perspectives. One way is to identify the winding modes of the family of type IIB strings on the circle with the KK modes of the torus; and the other way of looking is to identify KK modes of the circle with wrapping of the membrane on the torus. Again the mass formula matching relations can be used to relate parameters on both sides.

There is also a five brane, the soliton counter part of fundamental membrane, in the M-theory and these are the only two extended objects in the 11-dimensional theory. Therefore, one expects that M-theory should be able to give description of NS branes and Dp-branes in the lower dimensional string theories. Moreover, the membrane tension of 11-dimensional theory,  $T_2^{(M)}$ , is the only parameter since the 5-brane tension,  $T_5^{(M)}$  is determined from the Dirac quantization relation in terms of  $T_2^{(M)}$ . In order to study the branes in 9-dimensional theory, we should see how different branes arise from M-theory and from type IIB theory. The simplest is the 2-brane. In case of M-theory, the membrane remains a membrane; but type IIB in 10-dimensions has no membrane (due to the absence of even RR field strength), therefore, the D3-brane of ten dimensional IIB theory wraps around  $S^1$  to produce a membrane. But the type IIB string tension also is related to  $T_2^{(M)}$  since membrane wraps around torus to produce the string. So the simplest result is D3-brane tension and string tension of type IIB are related:  $T_3^{(B)} = \frac{1}{2\pi}(T_2^{(B)})^2$ ; this result involves only IIB theory tensions derived through M-theory route. The 9-dimensional IIB theory will have D3-branes too. They will arise, from M-theory view point, as wrapping of the 5-brane around  $T^2$ . There are 4-branes in 9-dimensional IIB theory. Since 10-dimensional IIB theory has  $SL(2, Z)$  pair of strings, their solitonic partners 5-branes will come in same multiplets too. These 5-branes, compactified on  $S^1$ , will give the 4-branes in 9-dimensions. Similarly, one can discuss type IIA theory from both the perspectives in ten dimensions. If we define  $L = 2\pi R_{11}$  as the perimeter of the compact circle, then tension of IIA string gets related to  $T_2^{(M)}$  since IIA string, in 10-dimensions, will arise due to wrapping of membrane around the circle. The relation is  $T_1^{(A)} = g_A^{-\frac{2}{3}}LT_2^{(M)}$ . For type IIA membrane we have  $T_2^{(A)} = g_A^{-1}T_2^{(A)}$ . The 4-brane in IIA theory will come from wrapping of M-theory 5-brane around the circle and the relation becomes  $T_4^{(A)} = g_A^{-\frac{5}{3}}LT_5^{(M)}$  and we must also have  $T_5^{(A)} = g_A^{-2}T_5^{(M)}$ . Then using relation between  $T_2^{(M)}$  and  $T_5^{(M)}$  together with the relation between  $T_4^{(A)}$  and  $T_5^{(A)}$ , one can get an expression:  $T_5^{(A)} = \frac{1}{2\pi}(T_2^{(A)})^2$ .

The purpose of above examples was to illustrate how one can derive a large number relations using the M-theory. In lower spacetime dimensions, the theory provides a rich basis to understand branes coming from various theories. We would like to record one important fact for our future considerations. Note that the formula (4.10) of general  $(q_1, q_2)$  string is for a string carrying NS and RR charges. For a string with only NS charge the  $(1, 0)$  the tension scales as  $T \sim g_B^{\frac{1}{2}}$  and for the one carrying only unit RR charge it is  $T \sim g_B^{-\frac{1}{2}}$ . In the string frame, after rescaling the metric, we find that a string with one unit of NS charge has tension of order 1 and the string with one unit of RR charge has tension  $g_B^{-1}$ . Therefore, the mass density also has same dependence on coupling constant.

There are duality relations which relate compactified M-theory to other string theories. One of the interesting cases is  $E_8 \times E_8$  heterotic string in ten dimensions [84]. There is no other string theory which can be related to this one. So it is expected that  $E_8 \times E_8$  is connected to the 11-dimensional theory. But it cannot be  $S^1$  compactified M-theory, because that compactified theory is type IIA as we have seen. Moreover, 11-dimensional theory as such is free from anomalies. However,

if one considers compactification on  $\frac{S^1}{Z_2}$  of M-theory it gets related to  $E_8 \times E_8$  ten dimensional theory. The orientation of  $S^1$  is reversed under  $Z_2$  and it flips the sign of 3-form potential  $C$ . As a consequence of this projection, we are left with the metric in ten dimension, the dilaton and 2-form potential. The gauge boson and 3-form potential  $C$  are projected out. The surviving fermions are Majorana-Weyl gravitino and Majorana-Weyl fermion. This is the supergravity in the bulk. Actually,  $\frac{S^1}{Z_2}$  is a line segment with fixed points at the boundary. These are two copies of 10-dimensional flat space. The states from twisted sector should be localised on these planes. It was shown that half of the anomalous variation is localized in one plane and the other half on the other plane. The possible gauge groups that can cancel the anomaly are  $E_8 \times E_8$  or  $U(1)^{496}$ . It is obvious the string theory to be identified is the  $E_8 \times E_8$  heterotic string. There are other duality conjectures [85, 86, 87] between M-theory and other string theories in lower dimensions: M-theory on  $K_3 \leftrightarrow heterotic/Type\ I\ on\ T^3$ . Compactification of M-theory on  $\frac{T^5}{Z_2}$  is dual to type IIB on  $K_3$ . The lower dimensional compactifications  $\frac{T^8}{Z_2}$  and  $\frac{T^9}{Z_2}$  are related to Type I/ heterotic on  $T^7$  and type IIB on  $\frac{T^8}{Z_2}$  respectively.

There are attempts to construct gauge supersymmetric gauge theories by choosing suitable combinations of intersecting branes and establish Seiberg-Witten dualities from this M-theory point of view of SUSY Yang-Mills gauge theories. Undoubtedly, the dualities together with the proposal for M-theory has brought us nearer to the goal of unified description of string theories. However, the underlying fundamental theory is yet to be discovered although we have seen many facets of that theory.

## 5 Black holes and String Theory

The physics of the black holes has many fascinating aspects. The classical black hole is the final stage of a collapsing heavy star. As the name suggests, matter falls into it and nothing comes out; there is an event horizon. However, deeper investigations have revealed, almost a quarter of a century ago, that there are strong similarities between thermodynamics and black hole mechanics [88, 89]. If  $M$  is mass of the black hole,

$$dM = \frac{1}{8\pi G} k dA, \quad \delta A \geq 0 \quad (5.1)$$

Here  $G$  is the Newton's constant,  $A$  is the area of the event horizon and  $k$  is the surface gravity. This is to be compared with thermodynamical relation,

$$dE = T dS, \quad \delta S \geq 0 \quad (5.2)$$

Hawking's startling discovery [90] that black holes radiate with a black body spectrum of temperature  $T = \frac{\hbar k}{2\pi}$ , when quantum effects are accounted for, raised several important issues in black hole physics. One can also associate entropy with a black hole

$$S_{BH} = \frac{A}{4G\hbar} \quad (5.3)$$

The thermodynamical relations used to describe macroscopic phenomena can be derived from statistical mechanics starting with microscopic fundamental laws of physics. Since  $\hbar$  appears in the black hole entropy formula, it is expected that the microscopic derivation of black hole entropy requires quantum gravity calculations. Moreover, entropy of a system, when interpreted from statistical mechanical point of view, counts the total number of degrees of freedom in the system. How do we count the number of degrees of freedom in a black hole and obtain the expression for entropy? There are more fundamental issues related to quantum mechanics when we carefully examine the implications of Hawking radiation. We can think of allowing some matter to go into the black hole, prepare the initial state as a pure quantum state to be the incident wave. However, the emitted Hawking radiation has a black body distribution and thus these are mixed states. Therefore, the S-matrix that will describe the above process will lose its unitarity property.

In the perturbative regime, string theory can provide reliable results for computations of processes involving graviton. The resulting S-matrix elements respect the required unitarity and analyticity properties. Thus, it is pertinent to ask what string theory has to offer in resolving the issues alluded to earlier. Recently, one of the important achievements of the string theory has been the microscopic derivation of the black hole entropy, for a special class of black holes that arise in string theory. We shall, initially, not set  $G = 1$ , to bring out a few salient points in discussions of stringy black holes and some times we shall display presence of  $\hbar$  in formulas. Recall, that the Newton's constant is related to string coupling and tension as  $G \sim g_{str}^2/T$ , in four spacetime dimensions. If we have a massive string state, the gravitational field is  $GM_s$ , where  $M_s$  is mass of a string state measured in units of  $T$ ; also some times we shall denote it as  $M$ . Thus, the field increases as string coupling increases. String states are given by the mass formula  $M^2 = NT$  and it is well known that at a given mass there are a lot of states and the degeneracy [92] grows exponentially with mass, i.e.  $e^M$ . Thus one might think that the excited states, if treated as black holes, will reproduce the entropy formula; however, this simple argument is not adequate since black hole entropy grows like  $M^2$ , whereas the naive argument will give  $S_{BH} \sim M$ . There have been attempts to explain this discrepancy saying that the mass that would appear in microscopic derivation of  $S_{BH}$  is not the same as the one appearing in Beckenstein-Hawking formula and there might be renormalization effects to be accounted for [91]. The perturbative string states appear in infinite levels and thus, for high enough mass, the massive elementary string state will lie inside the Schwarzschild radius associated with it. Consequently, they will require black hole descriptions. One of the ways to derive black hole entropy microscopically is to consider such BPS states, so that when string coupling gets strong, the state is unchanged. In this approach [93], first step is to pick up appropriate BPS state and compute the microscopic entropy. Next, compute the Beckenstein-Hawking entropy of the BPS state, it is also an extremal black hole, and verify whether the two ways of calculating entropy are in agreement. This is the first clue that string theory might explain black hole entropy in microscopic way. However, the black holes constructed from the elementary string states had some shortcomings while computing the entropy. The area of the event horizon, for such black holes, tends to zero as one approaches the extremal limit; moreover, the dilaton also diverges at the horizon in this limit. This problem was encountered for string states in the NS sector.

The D-brane in RR sector can come as elementary states and there are corresponding solitonic states contained in the full spectrum. We had argued in the context of type IIB  $SL(2, Z)$  strings that in string frame metric, NS states have tensions of order 1, whereas, D-strings had mass density of the order of  $\frac{1}{g_{str}}$ . For the solitons of NS sector the mass goes as  $\frac{1}{g_{str}^2}$ ; but the solitons for RR sector still have mass order  $\frac{1}{g_{str}}$ . In the weak coupling regime NS solitons and RR ones are heavy. We should account for the gravitational fields they produce, which is  $GM$ . In view of above discussions, (i) NS elementary states produce very low field and (ii) RR states also produce low field in weak coupling limit; field tends to 0 as  $g_{str} \rightarrow 0$ . We may argue that in this regime, flat spacetime is a good description of the geometry. Since we are dealing with BPS states, as string coupling increases the mass remains unchanged, but the gravitational field keeps increasing and after some critical coupling, the spacetime is not flat any more; we must employ general theory of relativity. If these states describe black holes, then we should be able to compute the degrees of freedoms associated with them. It is possible to construct black hole configuration such that the area of the horizon is not zero nor the dilaton diverges at the horizon, when we take the extremal limit. For five dimensional black holes, we need at least three charges to have nonzero area for the horizon together with constant value for the dilaton at the horizon. In case of the four dimensional black hole needs four charges in order to satisfy the requirement of nonzero horizon area and finite value of dilaton (at the horizon).

The black holes which we shall consider now have some special characteristics. They can be thought of as composites of many D-branes carrying Ramond charges. We have mentioned before that the BPS states have the property that mass of composite BPS state is the sum of the masses of the constituents. One starts in the weak string coupling phase with such D-branes and proceeds towards strong coupling domain when gravity becomes strong. In weak coupling regime, the

degeneracy of the level can be estimated reliably and microscopic entropy can be computed. In the strong coupling domain, the D-brane is inside the horizon and one can treat this like a black hole and compute the ratio  $\frac{A}{4G}$ , which is independent of string coupling  $g_{str}$  since both area and Newton's constant grow like  $g_{str}^2$ .

Let us discuss how the five dimensional black hole configuration is constructed with D-branes [94]. We start with type IIB theory in 10-dimensions. We know that it will admit D1-string and D5-brane. We want to make the composite object heavy; therefore, we put  $Q_5$  number of D5-branes and  $Q_1$  number of D1-strings together. Let us compactify this theory on  $T^5$  such that the  $Q_5$  number of D5-branes are wrapped around  $T^5$ , the  $Q_1$  D1-strings wrap along one of the directions of the torus. Then put some momentum along the direction in which the D-string wrapped; this momentum will be quantized in units of inverse radius of  $S^1$ . The aim is to evaluate the microscopic entropy by counting number of degrees of freedom for this system and it involves some detail technical steps [95, 96, 97, 98]; but we shall outline only essential points. We expect to have a  $U(Q_5)$  supersymmetric Yang-Mills theory on the D5-brane worldvolume. This will be a gauge theory in  $5 + 1$  dimensions which is derived by dimensional reduction of  $N = 1$  supersymmetric Yang-Mills theory from ten dimensions [95]. The D-string is inside this pack of D5-branes ( $Q_5$  of them). The D-string can be viewed as an instanton in this six dimensional spacetime, since an instanton in 6-dimensional theory with no time dependence and extension in one direction is a string. There are  $Q_1$  such strings in the D5-brane configuration. Their low energy dynamics is described by two dimensional supersymmetric sigma model in  $4Q_1Q_5$  dimensional hyper Kahler manifold. Every boson contributes factor 1 and every fermion contributes  $\frac{1}{2}$  to the central charge as we noted in Sec. II. Thus, total central charge is

$$c = 6Q_1Q_5 \quad (5.4)$$

Since we are dealing with BPS states, for these states  $L_0 = 0$  and the momentum given along  $S^1$  is related to the difference  $L_0 - \bar{L}_0$ . If we take momentum to be large i.e.  $P_s = -\frac{n}{R}$ ,  $n$  large; then using Cardy's result (relating degeneracy to central charge), one gets

$$d(Q_1, Q_5, n) = \exp(2\pi\sqrt{Q_1Q_5n}) \quad (5.5)$$

The black hole entropy computed from the microscopic view point is given by

$$S_{microscopic} = 2\pi\sqrt{Q_1Q_5n} \quad (5.6)$$

In order to derive the black hole entropy,  $S_{BH}$ , from Beckenstein-Hawking formula, we have to specify the metric, the charges and then compute the area of the event horizon in the extremal limit.

There is way to visualise the physical processes that lead to microscopic [99] derivation of the entropy formula. The D-string is inside D5-brane and the low level excitations are the lowest lying modes of the open strings attached to this one. If we think of the physical degrees of freedom, these are 8 transverse vectors and their super partners. Since these have to satisfy the Dirichlet boundary condition, they are constrained to move along the D-string. We are dealing with BPS state, therefore, these move only in one direction (say left). Since the D-string is wrapped around one circle of  $T^5$ , we choose  $x_1$ , then length is winding number times the radius of the circle. But the momenta of individual open strings moving on this unidirectional path on the circle is quantized. Moreover, sum of their momentum is constrained too by the total momentum we have put on that direction. Therefore, this is analogous to solving statistical mechanics of a one dimensional system on a circle where total energy (momenta are same as energy) is fixed.

The next step is to define the metric for the above configuration of the branes and the obtain the harmonic functions that are necessary to satisfy the equations of motion for the brane configurations [100, 101].

$$\begin{aligned} ds^2 = & H_1^{1/2} H_5^{1/2} \left\{ [H_1^{-1} H_5^{-1} (-K^{-1} dt^2 + K(dx_1 - (K^{-1} - 1)dt)^2) \right. \\ & \left. + H_5^{-1}(dx_2^2 + \cdots + dx_5^2) + dx_6^2 + \cdots + dx_9^2] \right\} \end{aligned} \quad (5.7)$$



We specify the compact directions as follows: the  $Q_5$  number of D5-branes are wrapped in  $x_1, \dots, x_5$  directions, D-string is wrapped in  $x_1$  and the momentum is along  $x_1$  too. Since we toroidally compactify to five dimensions  $x_i$ ,  $i = 1, \dots, 5$  are periodic and the radius of compactification is  $R_i$  along  $i$ th direction. and

$$e^{-2\phi} = H_1^{-1} H_5, \quad B_{01} = H_1^{-1} - 1 \quad (5.8)$$

$$H_{ijk} = \frac{1}{2} \epsilon_{ijkl} \partial_l H_5, \quad i, j, k, l = 6, \dots, 9 \quad (5.9)$$

$$r^2 = x_6^2 + \dots + x_9^2 \quad (5.10)$$

The harmonic functions are equal to

$$H_1 = 1 + C_1 \frac{Q_1}{r^2}, \quad C_1 = \frac{g_{str} \alpha'^3}{V} \quad (5.11)$$

$$H_5 = 1 + C_5 \frac{Q_5}{r^2}, \quad C_5 = g_{str} \alpha' \quad (5.12)$$

$$K = 1 + C_K \frac{Q_K}{r^2}, \quad C_K = \frac{n g_{str}^2 \alpha'^4}{R_1^2 V} \quad (5.13)$$

where  $V = R_2 R_3 R_4 R_5$ , we displayed the  $\alpha'$  dependence to show how the dimensionality of the charges appear, but now on we set the slope to unity as usual. Let us briefly note how the charges arise in this black hole. There is electric charge  $Q_1$  coming from  $B_{01}$  which is a gauge field now, after compactification of  $x_1$  coordinate. The  $Q_5$  charge is magnetic type originally attributed to D5-brane in 10-dimensions. After compactification the Poincare dual of that 3-form RR field strength is two form field strength and it becomes an electric charge counting D-brane charges. Of course, the third charge comes from momentum given along  $x_1$  direction and is quantized. When any one of these charges vanishes, the area of the event horizon vanishes too. The dimensional reduction [49] over the periodic coordinates  $x_1, \dots, x_5$ , yields the 5-dimensional effective action. The metric in the five dimensional space takes the following form

$$ds^2 = \lambda^{-2/3} dt^2 + \lambda^{1/3} (dr^2 + r^2 \Omega_3^2) \quad (5.14)$$

where

$$\lambda = H_1 H_5 K = (1 + C_1 \frac{Q_1}{r^2}) (1 + C_5 \frac{Q_5}{r^2}) (1 + C_K \frac{Q_K}{r^2}) \quad (5.15)$$

This corresponds to an extremal charged black hole and the horizon is located at  $r = 0$ . However, the area of the horizon is nonzero and it is proportional to the product of the charges. The expression for the area is

$$A_5 = (r^2 \lambda^{1/3})^{3/2} \Big|_{r=0} = \sqrt{C_1 Q_1 C_5 Q_5 C_K Q_K} (2\pi^2) = \frac{g_{str}^2}{R_1 V} \quad (5.16)$$

The Newton's constant in five dimensions gets related to the ten dimensional Newton's constant after we compactify on  $T^5$  and the relation is

$$G_N^{(5)} = \frac{G_N^{(10)}}{(2\pi)^5 R_1 V} = \frac{1}{4} \frac{\pi g_{str}^2}{R_1 V} \quad (5.17)$$

Therefore, the entropy is equal to

$$S_{BH} = \frac{A_5}{4G_5} = 2\pi \sqrt{Q_1 Q_5 n} \quad (5.18)$$

This expression exactly agrees with the expression for  $S_{microscopic}$ . A few comments are in order to discuss the constraints on the parameters for the above relation to be valid. The string effective action adopted to obtain the brane solutions is valid when string loop corrections and  $\alpha'$  corrections are nonleading. The string loop corrections are small when  $g_{str} \rightarrow 0$  with the values of the charges



held fixed. The charges correspond to characteristic scales of the system. If we want ignore  $\alpha'$  correction terms then the charges should be larger than string scale i.e.  $Q_1$ ,  $Q_5$  and  $n$  are much larger than  $\alpha'$ . If the compactification radii of the torii be taken as order of string length scale, then we should have  $g_{str}Q_1$ ,  $g_{str}Q_5$ ,  $g_{str}^2n \gg 1$ . This tells us that  $n \gg Q_1 \sim Q_5 \gg 1$ . The entropy of nonextremal black holes can be considered in a similar manner; however, we must keep several points in mind. First of all, the extremal black holes are BPS states and they get no quantum corrections. Therefore, whereas the microscopic entropy is computed in the weak coupling phase, the Beckenstein-Hawking entropy is obtained after we go over to the strong coupling domain so that the composite D-brane configuration lies inside the horizon. In case of nonextremal black holes, we have no theorem against quantum corrections and therefore, passage to strong coupling limit is not so simple. It is argued, that a black hole which is slightly away from extremality might allow smooth increase of the coupling constant as one starts from weak coupling limit. This type of black holes configuration can be achieved by allowing some low level right moving oscillators compared to the high left moving levels (note that for extremal case  $N_R = 0$ ). We shall not discuss the properties of these black hole in detail here.

The BPS extremal black holes are stable and they have zero temperature; therefore, they will not emit Hawking radiation. If we intend to understand the Hawking radiations from black holes in string theory, we have to look for those ones which are excited states and can decay into lower energy state. The starting point is to consider a nonextremal black hole. Since there will be left and right movers, the open string states will be going in opposite directions on the D-string. Again, it is a one dimensional problem where one can imagine that two oppositely moving open string states collide to give a closed string state. If we were to calculate the S-matrix element for such a process, we shall consider initial state, final state and a suitable interaction Hamiltonian for our computational purpose. In order to get the emission rate, one will take modulus square of this amplitude, average over initial states, sum over final states and divide by usual phase space factor. The state of the initial nonextremal black hole is given by occupation numbers  $N_L$  and  $N_R$  and the amount of momentum we give on the compact circle which are going in opposite directions. The momenta are quantized as  $\frac{n}{R}$  in either direction and thus the closed string state will carry momentum  $\frac{2n}{R}$ . As we have seen there are  $4Q_1Q_5$  bosonic and fermionic oscillators. The string theory calculation gives the amplitude for emission of a closed string state from these initial state [100]. The sum over final state and averaging over initial states leads to a factor  $\rho_L\rho_R$ , where for example

$$\rho_R = \frac{1}{N_i} \sum_i \langle i | N_R | i \rangle \quad (5.19)$$

where  $N_i$  is the total number of initial states and  $N_R$  is the number operator of right movers. We might carry out the averaging over all possible initial states with a given value of  $N_R$  by adopting the statistical mechanical prescription. The problem actually maps to the case of one dimensional gas and the microcanonical ensemble can be used since we are holding  $N_R$  fixed; energy is held constant. The configuration of the black hole is such that  $N_L \gg N_R > 1$ . If  $k_0$  is the momentum of out going massless closed string the final calculation give the decay rate as

$$d\Gamma \sim (\text{Area}) \frac{e^{-\frac{k_0}{T_R}}}{1 - e^{-\frac{k_0}{T_R}}} d^4k \quad (5.20)$$

A more careful calculation [102] reveals a surprising result that not only the form of thermal distribution is recovered, but also the numerical coefficients match with semi-classical results of Hawking. The result has been derived for four dimensional black holes as well [103]. It is an interesting question to ask whether one can calculate the absorption cross section of an extremal black hole for a closed string massless scalar and then relate that cross section to the decay rate of a nonextremal black hole by using the principle of detailed balance in quantum mechanics taking into account all the subtleties. Indeed explicit verification shows that such a check yields the correct result [104].

## 6 M-theory and the M(atrix) model

Our present understanding of string dynamics together with duality symmetries strengthen the belief that there is a fundamental theory and the five perturbatively consistent theories are different phases of that underlying theory. However, we do not know what this theory is except the conjecture that the low energy limit of this theory is the 11-dimensional supergravity action. There are deep questions about the structure of this theory. We shall call it the M-theory. We recall that strong coupling limit of the type IIA theory is identified with 11-dimensional supergravity. When viewed from type IIA perspective, the existence of D0-branes as nonperturbative RR point like objects is quite important for our discussion. They are BPS states and their mass is of the order  $\frac{1}{g_{str}}$  and scaled by 10-dimensional length scale  $l_s$ . These being BPS states, one could assume that there are threshold bound states of many, say  $N$ , D0-branes which satisfy the properties of bound BPS states. Now if we take the strong coupling limit, then it is found that the low energy spectrum is same as the spectrum of the 11-dimensional supergravity. This is an important evidence. Furthermore, the 11-dimensional theory is known to admit membrane and five brane and we have argued how one can study properties of various brane configurations in string theories after compactifying the M-theory. The M(atrix) model [105, 32, 33] can describe perturbation expansions of various string theories. There is a limit in which the theory provides connection with 11-dimensional supergravity theory. However, one would like to seek answers to several questions from this theory. For example, the general prescription for the compactification of the theory is not known. Similarly, the complete set of degrees of freedom of this theory is to be obtained. The M(atrix) theory, nevertheless, provides insight into nonperturbative definition of string theory and it also exhibits string dualities [106]. One can also go over to various string theories by adopting different limiting prescriptions.

The model resorts to infinite momentum frame (IMF) technique boosted along a compact direction. The momenta along compact direction is quantized; and one starts with  $N$  units of these momenta and then  $N \rightarrow \infty$  limit is taken. Since one is working in the light-cone frame while constructing M(atrix) theory, the theory is not manifestly Lorentz invariance. Thus Lorentz invariance might be recovered in the large  $N$  limit. In the M(atrix) model formulations one encounters parameters which have the interpretation of being expectation values of scalars when viewed from the string theory side. But in the M(atrix) model when we have IMF formulation, these constant modes have infinite frequency and they are frozen into fixed configuration. The theory in its present formulation is not background independent. Moreover, one encounters problems while compactifying the theory on an arbitrary  $d$ -dimensional torus. We may remind the reader that the M(atrix) theory provides a rich structure to study various aspects of string theory from M-theory stand point.

The infinite momentum frame (IMF) technique played a very useful role in current algebra [107]. In field theoretic calculations it simplifies perturbation theory calculations [108, 109]. When we have to deal with a collection of particles, we can define IMF to be a frame where the total momentum is taken to be very large. If we designate particles by index  $I, J, \dots$  then

$$P_I = \eta_I P + P_{TI} \quad (6.1)$$

where  $T$  stands for 'transverse' and  $P \cdot P_{TI} = 0, \sum P_{TI} = 0$  and  $\eta_I \leq 1$ . For a highly boosted coordinate system we could have all  $\eta_I$  positive. Particularly, for the case at hand, we deal with massive particles and we can choose an appropriate frame to satisfy our requirement. Energy of any particle satisfies relativistic relation

$$E_I = \sqrt{P_I^2 + m_I^2} = \eta_I P + \frac{P_{TI}^2 + m_I^2}{2\eta_I P} + \dots \quad (6.2)$$

it is understood that there are terms higher order in  $\frac{1}{P}$  denoted by dots. The expression for energy is similar to that of a nonrelativistic particle in a lower dimension with mass term taking a modified form. When we use a light-cone (LC) frame, a spatial direction is identified and designated as longitudinal. The longitudinal momentum is  $P_{LI} = \eta_I P$  and one defines  $P_{\pm I} = E_I \pm P_{LI} = E_I \pm \eta_I P$ . The mass shell condition translates to  $P_{+I}P_{-I} - P_{TI}^2 = m_I^2$  and we can rewrite this

relation as

$$E_I - \eta_I P = \frac{P_{TI}^2 + m_I^2}{P_{I+}} \quad (6.3)$$

In the limit of large  $P$ , we have  $\eta_I P$  large and therefore,  $E_I \rightarrow \eta_I P$  with  $P_{I+} \sim 2\eta_I P$ . When M-theory is envisaged in IMF, let us designate the momenta as  $p_0, p_i, i = 1, \dots, 9$  and  $p_{11}$ . One compactifies 11th direction with and this is also boosted, therefore  $\{p_i\}$  are collectively denoted as  $p_T$ . Thus for collection of the D0 particles

$$E - p_{11}^{total} = \sum_I \frac{p_{TI}^2}{2p_{I11}} \quad (6.4)$$

We note that there are 32 real supercharges in the theory. When one adopts IMF description, it is convenient to split them into two groups each having 16 of them. The charges in every group transform as spinors of  $SO(9)$ . Let us denote charges as  $Q_\alpha, \alpha = 1, \dots, 16$ , and  $q_A, A = 1, \dots, 16$ . The algebra of these charges are

$$\{Q_\alpha, Q_\beta\} = \delta_{\alpha\beta} H, \quad \{q_A, q_B\} = \delta_{AB} P_{11} \quad (6.5)$$

$$\{Q_\alpha, q_A\} = \gamma_{A\alpha}^i P_i \quad (6.6)$$

Here  $H$  is the Hamiltonian operator,  $P$ 's are the corresponding momentum operators and  $\gamma_i$  are 16 dimensional gamma matrices.

We have discussed earlier, how D0-brane has a natural interpretation from the 11-dimensional theory with a compact coordinate and the RR charge is related to quantized momenta along this direction. The relation between mass and charge is satisfied since these are BPS states. There exists a sector with  $N$  units of D0-brane charge, carrying Kaluza-Klein momentum  $\frac{N}{R_{11}}$ . If we hold  $N$  fixed and take a limit  $R_{11} \rightarrow 0$ , we go over to the weak coupling phase of string theory; however, in the passage to this limit, the string scale is not held fixed. The aim is to study the phenomena in the 11-dimensional theory and thus  $l_{11}$  is to be kept fixed. We recall that

$$R_{11}^3 = g_{str}^2 l_{11}^3 \quad (6.7)$$

and the string length scale,  $l_s^2 = \frac{l_{11}^2}{R_{11}}$ . Thus as the compactification radius tends to zero string scale diverges. We have also seen earlier, as the radius shrinks, the mass of D0-brane tends to infinity, when measured in 11-dimensional Planck units in ten dimensions. In other words the mass of D0-brane is

$$\frac{1}{g_s l_s} = \frac{1}{R_{11}} \quad (6.8)$$

and therefore, it is appropriate to identify them as the KK modes. Thus, when we consider mass of these particles in 10-dimensions, in scales of eleven dimensional theory, the particles become very heavy and a nonrelativistic description is quite adequate. If we were to describe M-theory in terms of type IIA zero branes, then we have a scenario where M-theory is equivalent to  $N \rightarrow \infty$  limit of the nonrelativistic quantum mechanics of  $N$  D0-branes which are in weak coupling phase of type IIA theory. Furthermore, as Witten has argued [95] the physics of  $N$  coincident D0-branes is described by dimensionally reducing ten dimensional  $U(N)$  supersymmetric Yang-Mills theory to 0+1 dimensions [110]. Let us consider supersymmetric quantum mechanics of a single D0 particle. The starting point is the action

$$\int dt \text{Tr} \left( \frac{1}{2g_{str}} (D_0 X^i)^2 - i\theta^T D_0 \theta + \frac{1}{4g_{str}} ([X^i, X^j])^2 + \theta^T \gamma_i [X^i, \theta] \right) \quad (6.9)$$

This is the action obtained from 10-dimensional super Yang-Mills theory reduced to one dimension. Here  $i = 1, \dots, 9$  stands for transverse directions and  $\theta$  are real spinors with 16 components. Since  $X^i$  and  $\theta$  come from the gauge groups, they are in the adjoint representations of  $U(N)$ . Since they carry only time dependence, these are  $N \times N$  matrices.  $D_0 = \partial_t + [A_0, \cdot]$  is the covariant derivative and this can be converted to ordinary derivative with the gauge choice  $A_0 = 0$ . The mass of

D0-brane is order  $\frac{1}{g_{str}}$ , thus the first term in (6.9) can be written as  $\int dt \frac{M}{2} (\frac{dX^i}{dt})^2$ . Note that the action (6.9) contains parameters of type IIA theory. It is convenient to scale  $X^i = g_{str}^{\frac{1}{3}} Y^i$  which amounts to rescaling of the metric to that of 11-dimensional theory. Moreover, one scales the time variable as  $t = g_{str}^{\frac{2}{3}} \tau$  and denotes the  $\tau$  derivative by a dot. The action is rewritten as

$$S = \int d\tau \text{Tr} \left( \frac{1}{2R_{11}} (\dot{Y}^i)^2 - i\theta^T \dot{\theta} + \frac{R_{11}}{4} ([Y^i, Y^j])^2 + R\theta^T \gamma_i [Y^i, \theta] \right) \quad (6.10)$$

If  $\Pi_i = \frac{\dot{Y}^i}{R_{11}}$  and  $\pi = -i\theta^T$  are conjugate momenta of  $Y^i$  and  $\theta$  respectively, the corresponding Hamiltonian is given by

$$H = R_{11} \text{Tr} \left( \frac{1}{2} \Pi_i^2 - \frac{1}{4} ([Y^i, Y^j])^2 - \theta^T \gamma_i [Y^i, \theta] \right) \quad (6.11)$$

One can define  $H \equiv R_{11} \bar{H}$  for convenience factoring out over all  $R_{11}$ . Notice also that the potential energy term  $\frac{1}{4} R_{11} \text{Tr} ([Y^i, Y^j])^2$  is non-negative. When  $R_{11} \rightarrow \infty$ , we are in decompactification phase of M-theory. Thus, the finite energy states of H are those for which the Hamiltonian  $\bar{H}$  has vanishing eigenvalues. One seeks those states for which  $\bar{H}|\psi\rangle = \frac{\epsilon}{N} |\psi\rangle$ , which is equivalent to seeking a solution  $H|\psi\rangle = \frac{R_{11}}{N} |\psi\rangle$  where  $\epsilon$  is finite. We know that, for collection of N number of D0-branes, the total momentum  $p_{11} = \frac{N}{R_{11}}$  and therefore, the energy is given by  $E = \frac{\epsilon}{p_{11}}$ . We have to identify  $\epsilon$  with  $\frac{1}{2} P_T^2$  if we recall (6.4). The  $N \times N$  matrices  $X^i$  can be interpreted as the location of N D0-branes. When we consider the potential term in  $Y^i$  variables (6.10), we notice that there are flat directions when  $[Y^i, Y^j] = 0$ . Here we deal with a quantum mechanical system and  $Y^i$  have are the collective coordinates. In such situation as mutually commuting  $Y^i$ , we can diagonalize  $Y^i = \text{diag} (y_1^i, y_2^i \dots y_N^i)$ . Thus  $y_n^i$  is the  $i$ th coordinate of the  $n$ th D0-brane. It is easy to see that there is invariance under Galilean translation,  $Y^i \rightarrow Y^i + d^i \mathbf{1}$  and the Galilean boost  $Y^i \rightarrow Y^i + v^i t \mathbf{1}$  as is expected of a nonrelativistic system, here  $\mathbf{1}$  is the unit matrix. The boost will affect the center of mass momentum; but neither the relative momenta nor interaction term are affected by these transformations.

We can consider two clusters separated from one another. This is familiar in composite model of hadrons where quarks are the basic constituents. In the parton picture, the proton is made of large number of partons with very small binding energy and one could describe photon-hadron deep inelastic scattering in IMF [111]. In this case we can think of configurations where the  $N \times N$  matrices  $Y^i$  can be decomposed to block diagonal form of say n blocks of  $N_1 \times N_1, N_2 \times N_2, \dots, N_n \times N_n$  such that  $\sum_m N_m = N$ . This decomposition can be interpreted as if we have n separated clusters of D0-branes where each of the clusters has  $N_1, N_2, \dots, N_n$  number of particles. The distance between two clusters can be defined as

$$r_{ab} = \left| \frac{1}{N_a} \text{Tr} Y_a - \frac{1}{N_b} \text{Tr} Y_b \right| \quad (6.12)$$

where  $a$  and  $b$  are the two clusters. Now we can visualize how the potential will arise. It comes from  $\text{Tr} ([Y^i, Y^j])^2$  and this goes like modulus squared of the off diagonal block elements multiplied by the minimum of the  $r_{ab}^2$  and an appropriate numerical constant. Thus, if we consider well separated cluster of D particles, the off diagonal elements are required to be small; otherwise, the potential will grow like  $r_{ab}^2$ . We should keep in mind that the system is supersymmetric and having a harmonic oscillator type potential does not imply ground state energy is that of the oscillator. The supersymmetric quantum mechanical system has a very rich structure. This becomes transparent if we consider a single D0-brane, i.e.  $N = 1$ .

$$H = \frac{R_{11}}{2} \Pi_i^2 \equiv \frac{R_{11}}{2} P_T^2 = \frac{P_T}{p_{11}} \quad (6.13)$$

When we look at this equation from 11-dimensional point of view, this corresponds to the relation between energy and momentum of a massless particle in IMF. When we take into account the

16 component fermions,  $\theta$  we eventually get the supermultiplet with 256 total degrees of freedom and this agrees with the massless degrees of freedom of  $N = 1$  supergravity in eleven dimensions. In fact the bosonic components are 128 equal to fermionic degrees of freedom. As is well known, there 44 components from graviton and 84 from the antisymmetric tensor field in 11-dimensions. When we have  $N > 1$ , it is necessary to separate the center of mass motion and define the relative coordinates and the decomposition is as follows:

$$Y^i = Y_r^i + Y_{cm}^i \mathbf{1}, \quad Y_{cm}^i = \frac{1}{N} \text{Tr } Y^i \quad (6.14)$$

$$\Pi_i = \Pi_{r\ i} + \frac{1}{N} P_{cm\ i} \mathbf{1}, \quad P_{cm\ i} = \text{Tr } \Pi_i \quad (6.15)$$

and  $\text{Tr } Y_r^i = \text{Tr } \Pi_{r\ i} = 0$ . Now the total Hamiltonian will be written as a sum of two terms

$$H = H_{cm} + H_r \quad (6.16)$$

with

$$H_{cm} = \frac{R_{11}}{2N} (P_{cm\ i})^2 = \frac{1}{p_{11}} (P_{cm\ i})^2 \quad (6.17)$$

Note the appearance of the factor  $\frac{R}{N} = \frac{1}{p_{11}}$  as expected. We have defined the center of mass coordinate, canonical momentum and the Hamiltonian by taking trace over  $U(N)$  matrices. Therefore, the relative Hamiltonian is a function of  $\{Y_r^i, \Pi_{r\ i}\}$ . Thus  $H_r$  is quite similar to the original Hamiltonian; however, all the variables are  $SU(N)$  matrices, they are traceless since the trace part is separated out. It has been shown that the relative Hamiltonian has zero energy bound states due to the presence of supersymmetry [95, 112, 113]. The total energy is due to the center of mass energy:  $E = E_{cm} = \frac{1}{2p_{11}} (P_{T\ cm})^2$ . In this case one also gets the supergravity multiple which has 256 states. Therefore, for any  $N$ , we see that the spectrum contains supergravitons. Suppose we decompose  $Y^i$  to various blocks which describe clusters of D-particles. In the simplest case, if the submatrices are exactly block diagonal so that off diagonal elements are zero, then the total Hamiltonian will be given by sum of  $n$  separate Hamiltonians without any interactions amongst them. If we let the off diagonal elements appear (give them small values), that will amount to switching on interactions between the clusters. The physical picture is that we have several clusters, each cluster will have its supergraviton in the spectrum. There could be arbitrary number of them and therefore, we let  $N$  go to infinity. Thus the matrix model contains the full Fock space of supergravitons. The interaction among the supergravitons is described due to the presence of off diagonal elements and one should be able to describe various processes involving supergravitons in this picture.

In order to compute S-matrix element for scattering of two supergravitons when their transverse velocities are small, we have to determine potential between them. One starts by considering the classical configurations and the fluctuations over them to compute the effective action [114]. Suppose we give transverse velocity  $v$  and define the impact parameter as  $b$  and expand the coordinates around their backgrounds as follows:

$$X^9 = \frac{1}{2} b \sigma_3 + \sqrt{g_{str}} \delta X^9, \quad X^8 = \frac{1}{2} v t \sigma_3 + \sqrt{g_{str}} \delta X^8 \quad (6.18)$$

$$X^i = \sqrt{g_{str}} \delta X^i, \quad i \neq 8, 9 \quad (6.19)$$

Here  $\delta X^i$  denotes the fluctuations and  $\sigma_3$  is the Pauli matrix. When we have vanishing fluctuation, the classical configuration is such that total transverse center of mass momentum and position vanish. The  $2 \times 2$  matrices are block diagonal which describes two clusters of D0-branes and in this case we have  $N_1 = N_2 = 1$ . Now the separation between the two particles is given by  $r_{ab} = \sqrt{v^2 t^2 + b^2}$ . The effective action can be computed using the standard techniques and the order  $\hbar$  term will contain determinant of (basically) propagators when we restrict to one loop level. Thus

$$S_{eff} = S_0 + \int d\tau V_{eff}(r(\tau)) \equiv \int d\tau V_{eff}(\sqrt{v^2 \tau^2 + b^2}) \quad (6.20)$$

For large impact parameter, the long range part of the potential in the leading order is given by [114]

$$V_{eff}(r) = -\frac{15}{16} \frac{v^4}{r^7} + \text{higher orders} \quad (6.21)$$

The result is striking in the sense that this form of the potential can be derived from the supergravity action at the tree level i.e. considering graviton exchange. Thus starting from a simple M(atric) model description, one could extract a result of 11-dimensional supergravity.

The 11-dimensional supergravity admits supermembrane. It is worthwhile to ask how much the M(atric) model can tell us about the underlying membrane theory. The membrane is extended object in two spatial directions as the name suggests. Moreover, the dimension of spacetime in which the supermembrane can exist is quite restricted [115, 25]. The reason for such constraints lies in the fact that the action contains Wess-Zumino-Witten term and the supersymmetry invariance of the full action restricts the spacetime dimensions to 4,5,7 or 11. The membrane is described by  $Z^\mu(\sigma, \xi, \tau)$ , where  $\sigma, \xi$  and  $\tau$  are the worldvolume coordinates. When one adopts a Hamiltonian formalism, a fixed  $\tau$ -slice is chosen and thus the explicit  $\tau$  dependence in  $Z^\mu$  does not appear and the derivatives with respect to worldvolume time are traded for canonical momenta  $\mathcal{P}_\mu$ . The light-cone gauge is a convenient description to see the physical degrees of freedom and in this gauge the membrane Hamiltonian takes the following form [116]

$$H_M = \frac{1}{2p_{11}} \int \frac{d\sigma d\xi}{(2\pi)^2} \mathcal{P}_i^2 + \frac{(2\pi T_2)^2}{4p_{11}} \int d\sigma d\xi (\{Z^i(\sigma, \xi), Z^j(\sigma, \xi)\})^2 + \text{fermionic terms} \quad (6.22)$$

where the brackets appearing in the second term are defined as

$$\{A, B\} = \partial_\sigma A \partial_\xi B - \partial_\xi A \partial_\sigma B \quad (6.23)$$

and  $T_2$  is the membrane tension. Let us assume that the worldvolume of the membrane can be written as  $\Sigma \times R$ , where  $\Sigma$  has the topology of a torus. For this topology,  $Z^i(\sigma, \xi)$  is a double periodic function and we can expand  $Z^i$  in double Fourier series with  $Z_{mn}^i$  as the Fourier coefficients. Thus we have nine  $\infty \times \infty$  matrices and same would be the case if we had considered nine  $Y^i$ 's in the  $N \rightarrow \infty$  limit. In order to establish relation with the membrane Hamiltonian (6.22), we have show how the commutator  $[Y^i, Y^j]$  goes over to the bracket  $\{Z^i, Z^j\}$ . For arbitrary finite  $N$ , introduce two  $N \times N$  matrices  $U$  and  $V$ , satisfying the properties

$$U^N = V^N = 1, \text{ and } UV = e^{\frac{2i\pi}{N}} VU \quad (6.24)$$

This can be realized if  $U$  and  $V$  have the following special form  $U_{j,j+1} = U_{N,1} = 1$  and  $V_{j,j} = e^{\frac{2i\pi(j-1)}{N}}$  and all other matrix elements set to zero. A more abstract, 't Hooft, representation is

$$U = e^{ip}, \quad V = e^{iq}, \quad [p, q] = \frac{2\pi}{N} i \quad (6.25)$$

This is the canonical commutation relation between position and momentum when the space is taken to be compact and discrete. It is worthwhile to point out that the above commutation relation will not hold good for finite dimensional matrices. However, acting on states with low wave number, the error on the r.h.s of the commutator  $[p, q]$  is further down by power of  $N$  and therefore,  $\frac{2\pi}{N}$  is the leading term. Thus as  $N$  assumes higher and higher values, the error gets smaller and smaller. From  $U^N = V^N = 1$  we can conclude that  $p$  and  $q$  take eigenvalues  $m \frac{2\pi}{N}$ , where  $m$  takes values  $0, 1, 2, \dots, (N-1)$ . Moreover  $\text{Tr} U^n V^m = N \delta_{n,0} \delta_{m,0}$ , where 0 in both the Kronecker delta are to be understood as *mod*  $N$ . Now we can expand any  $N \times N$  matrix in terms of Fourier modes.

$$A = \sum_{n,m=N/2-1}^{N/2} A_{nm} U^n V^m = \sum_{n,m=N/2-1}^{N/2} A_{nm} e^{in p} e^{im q} \quad (6.26)$$

Since commutator of  $p$  and  $q$  is order  $\frac{1}{N}$ , in the  $N \rightarrow \infty$  limit, they will commute. The eigenvalues of these two operators will fill the interval  $[0, 2\pi]$  and 0 is to be identified with  $2\pi$  since we have

toric geometry. The double Fourier expansion (6.26) takes the form

$$A(p, q) = \sum_{n, m=-\infty}^{\infty} A_{nm} e^{inp} e^{imq} \quad (6.27)$$

and the Fourier coefficients with the double index are defined as

$$A_{nm} = \int_0^{2\pi} \int_0^{2\pi} \frac{dp}{2\pi} \frac{dq}{2\pi} A(p, q) e^{-inp} e^{-imq} \quad (6.28)$$

Also  $\text{Tr } A = NA_{00}$ , when we take  $N \rightarrow \infty$  limit,  $\text{Tr } A \rightarrow N \int_0^{2\pi} \int_0^{2\pi} \frac{dp}{2\pi} \frac{dq}{2\pi} A(p, q)$ . One can show with some algebra that the commutator of two matrices in the infinite  $N$  limit goes over to the  $\{, \}$ . Finally bosonic part of the M(atrrix) model Lagrangian goes over to a form (identify  $\frac{dp}{2\pi} = d\sigma$  and  $\frac{dq}{2\pi} = d\xi$ )

$$L_m \rightarrow \frac{N}{2R} \int d\sigma d\xi (\dot{Z}^i(\sigma, \xi))^2 - \frac{R}{4N} \int d\sigma d\xi (Z^i(\sigma, \xi), Z^j(\sigma, \xi))^2 \quad (6.29)$$

Note that  $\frac{N}{R} = p_{11}$ , therefore conjugate momentum of  $Z^i$  is  $p_{11} \dot{Z}^i$ . Thus passage to the Hamiltonian (in light-cone gauge) gives the membrane Hamiltonian (6.22). This is indeed a remarkable result that a simple supersymmetric quantum mechanical system encodes the dynamics of the supermembrane.

It is natural to ask whether one obtain a string starting from the M(atrrix) model. First, one compactifies the theory to ten dimension. When the compactification radius is small, the theory contains the Fock space of the type IIA string. As the radius tends to zero the string becomes free [117] and correct leading order string interactions could be reproduced. In order to carry out compactification, we replace the matrices by infinite dimensional operators. The compact coordinate is represented as  $X^a \rightarrow -i \frac{\partial}{\partial \sigma^a} \mathbf{I}_{N \times N} - A_a(\sigma)$ . Here  $A$  is a  $U(N)$  gauge potential. The rest of the variables are taken to be matrix valued function of  $\sigma$ . If we use this ansatz, the resulting Hamiltonian is that of maximally supersymmetric 1+1 dimensional Yang-Mills theory. In the limit when radius goes to zero and  $N$  is taken to be infinity, the moduli space of this model coincides with the Fock space of type IIA theory.

Indeed, the M(atrrix) model has opened up new avenues to study dualities between compactified model on torus and Yang-Mills theory on dual space. Moreover, there are applications of the M(atrrix) model to study black holes we refer the interested reader to the review on the subject [33]. Another interesting development has been to understand type IIB theory and its dualities from a matrix model formulation. In this approach one adopts procedure of Eguchi and Kawai to consider reduced 10-dimensional super Yang-Mills theory and it is a theory of  $N \times N$  matrices which even carry no time dependence [118]. We refer the reader to the review article of Makeenko [119].

## 7 Anti-de Sitter Space and Boundary field Theory Correspondence

Recently, attentions have been focused in constructing supersymmetric gauge theories by considering various configurations of branes in string theories as well as in M-theory. When we have  $N$  coincident Dp-branes, a supersymmetric  $U(N)$  gauge theory lives in worldvolume of the branes. The  $\frac{1}{N}$  expansion proposed by 't Hooft [121] revealed several aspects of  $SU(N)$  Yang-Mills theory. According to 't Hooft, one should consider large  $N$  limit of the theory keeping  $g_{YM}^2 N$  fixed,  $g_{YM}$  being the gauge coupling constant. Then a Feynman diagram is designated by the topological factor  $N^\chi$ ,  $\chi$  being the Euler characteristic of the Feynman diagram. When we consider, expansion in  $\frac{1}{N}$ , rather than in coupling constant, each order in  $\frac{1}{N}$ , contains diagrams to all orders in coupling constant and the leading order corresponds to the planar diagrams. Maldacena [120] has made remarkable conjecture regarding large  $N$  conformal gauge theories. The proposal states that large

N limit of a conformally invariant theory in  $d$  dimensions is determined by supergravity theory on  $d + 1$  dimensional Anti-de Sitter space times a compact space (for a sphere it is maximally supersymmetric). The AdS/CFT connection has led to the generalization of the holography principle in this context [123, 124] which was first introduced in black hole physics [125, 126] in order to understand the Bekenstein entropy bound and the area law for black hole entropy. Thus the conjectures of Maldacena led to reveal deeper connections between string theory and superconformal gauge theories.

We have emphasized earlier that gravity is an integral part of string theory since graviton is a part of the spectrum. Moreover, gauge fields also invariably appear in string theories. Let us recapitulate a few points in order to get a perspective of AdS/CFT connections. We have seen that the heterotic strings, through their constructions, contain nonabelian gauge groups and graviton in their massless spectrum. The type II theories have graviton, coming from NS sector, in their perturbative spectrum. However, with the discovery of Dp-branes, we know that supersymmetric gauge theories can arise if we consider coincident Dp-branes in type II theories. Type I string theory admit nonabelian gauge field since Chan-Paton factors can be attached to the end points as was discussed earlier. Furthermore, consistency of the theory requires that we have to incorporate closed string sector in order to account for nonplanar loop corrections; therefore there is gravity coming from the closed string spectrum. For this theory, when we take  $\alpha' \rightarrow 0$  limit Yang-Mills theory appears automatically and since consistency requires inclusion of closed string states, gravity also will appear in the zero slope limit. In view of preceding remarks, one might conclude that, in string theory, gravity and gauge theory invariably appear simultaneously. Thus the important question to answer is that how the string theory can describe the strong interaction among quarks and gluons. The recent developments [120, 122, 123, 127] have provided connections between string theory and gauge theories.

The configuration under consideration is N coincident Dp-branes and open strings can end on these hypersurfaces. When we look into the dynamics in the worldvolume we have collection of these open strings and their excitations. Moreover, the worldvolume fields have their interactions and also there exists interaction with the bulk. An interesting limit to consider is when dilaton remains at a fixed value and the slope parameter tends to zero value. Then, at low energies, the gravity decouples; but to keep the interactions in the worldvolume in tact, we should have gauge coupling finite, for the  $U(N)$  gauge theory. In fact, if we ignore the center of mass part, then we need to consider the  $SU(N)$  gauge theory. It is necessary to go near the horizon,  $r \rightarrow 0$ , to see the connection between AdS and CFT. In the near horizon limit, recall eq.(5.11) and eq.(5.12), that the factor 1 appearing in the definition of the harmonic function of the Dp-brane can be neglected. To be specific let us first consider the metric in the case of N coincident branes.

$$ds^2 = H_p^{-\frac{1}{2}}(r)\eta_{\mu\nu}dx^\mu dx^\nu + H_p^{\frac{1}{2}}dy_idy_i \quad (7.1)$$

where,  $\{y_i\}$  are the transverse coordinates and  $r = \sqrt{y_i y^i}$ . The indices  $\mu, \nu, \dots$  are for tensors on the worldvolume. The dilaton and the  $(p + 1)$ -form potential, coming from the RR sector, are given by

$$e^{-(\phi - \phi_0)} = H_p(r)^{\frac{(p-3)}{4}}, \quad \text{and} \quad A = [H_p(r)]^{-1} \quad (7.2)$$

and

$$H_p(r) = 1 + \frac{C_p N}{r^{7-p}}, \quad C_p = \frac{(2\pi\sqrt{\alpha'})^{7-p}}{(7-p)\Omega_{8-p}} g_{str} \quad (7.3)$$

Here we have suppressed the indices of the  $(p + 1)$ -form gauge potential and  $\Omega_r = \frac{2\pi^{\frac{(q+1)}{2}}}{\Gamma[\frac{(q+1)}{2}]}$  and  $\phi_0$  is the asymptotic constant value of the dilaton. When we have N coincident D-branes, the worldvolume action is the generalised Born-Infeld action proposed by Tseytline [128]

$$S_{BI} = -\tau_p^{(0)} \int d^{p+1}\xi e^{-\phi} \text{STr} \sqrt{-\det[G_{\mu\nu} + 2\pi\alpha' F_{\mu\nu}]} \quad (7.4)$$



Here  $G_{\mu\nu}$  is the pullback of the metric  $G_{MN}$  to the world volume and  $F_{\mu\nu}$  is the gauge field strength on the brane. The tension of the brane is

$$T_p = \frac{(2\pi\sqrt{\alpha'})^{(1-p)}}{2\pi\alpha'g_{str}} = \frac{\tau_p^0}{g_s} \quad (7.5)$$

and  $g_{str}$  is the string coupling constant. The action (7.4) under the square root can be expanded and keeping the second order term in gauge field strength one can write the action in more familiar form

$$S_{gauge} = -\frac{1}{4g_{YM}^2} \int d^{p+1}\xi \text{Tr} F_{\mu\nu} F^{\mu\nu} \quad (7.6)$$

where Tr is taken over the gauge group matrices and the gauge coupling constant is identified as  $g_{YM}^2 = 2g_{str}(2\pi)^{(p-2)}(\alpha')^{\frac{(p-3)}{2}}$ . We know from the solutions discussed in previous section (recall eq.(5.12) and eq.(5.13)) that, in the limit, when  $r \rightarrow \infty$ , the metric is flat. Here one is looking for the behaviour of the solution in the  $r \rightarrow 0$  limit and one chooses a brane for which the dilaton is constant at the horizon. If we consider D3-branes, then we find that not only the dilaton is independent of  $r$ , but also the Yang-Mills coupling constant is dimensionless. As mentioned above, one examines the configuration of  $N$  coincident branes in the following limit

$$r \rightarrow 0, \quad \alpha' \rightarrow 0, \quad \text{and} \quad U \equiv \frac{r}{\alpha'} = \text{fixed} \quad (7.7)$$

Therefore, we can neglect 1 appearing in the harmonic function. and the D3-brane metric goes over to

$$\frac{ds^2}{\alpha'} \rightarrow \frac{U^2}{\sqrt{4\pi N g_{str}}} (dx_{3+1})^2 + \frac{\sqrt{4\pi N g_{str}}}{U^2} dU^2 + \sqrt{4\pi N g_{str}} d\Omega_5^2 \quad (7.8)$$

The last term is the line element of five sphere and the metric describes the manifold  $AdS \times S_5$ . The radius of AdS is the same as that of  $S_5$  and the radius is given by  $R_{AdS} = (\alpha'\sqrt{4\pi N g_{str}})^{\frac{1}{2}}$ . Since the Yang-Mills coupling constant satisfies the relation  $g_{YM}^2 = 4\pi g_{str}$ , the radius of the AdS gets related to the Yang-Mills coupling constant as

$$\frac{R_{AdS}^2}{\alpha'} = \sqrt{N g_{YM}^2} \quad (7.9)$$

We know that the worldvolume theory of  $N$  coincident Dp-branes is supersymmetric Yang-Mills theory in  $p+1$  dimensions and therefore, in this case the  $N=4$  SUSY gauge theory will appear. This is known to be a conformally invariant theory. From the supergravity side, we could describe the theory even for large radius; but that will amount to taking  $N g_{YM}^2$  to large values. Maldacena's conjecture states that strongly coupled  $N=4$  super Yang-Mills theory is equivalent to 10-dimensional supergravity compactified on  $AdS_5 \times S_5$ . However, the consistency of the supergravity theory requires string theory at a deeper level. Thus supersymmetric four dimensional Yang-Mills theory is equivalent to type IIB theory compactified on  $AdS_5 \times S_5$ . The relations among the parameters are

$$g_{YM}^2 \equiv \frac{\lambda}{N} = 4\pi g_{str}, \quad \text{and} \quad R_{AdS}^2 = \alpha' \sqrt{\lambda} \quad (7.10)$$

Let us very briefly recall some essential features of the Anti-de Sitter space. The Einstein-Hilbert action in the presence of cosmological constant term is

$$S_{EH} = \frac{1}{16\pi G_D} \int d^D x \sqrt{|g|} [R + \Lambda] \quad (7.11)$$

We consider D-dimensional spacetime with Minkowski metric. The field equations are

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = \frac{1}{2} g_{\mu\nu} \Lambda \quad (7.12)$$

Taking the trace of this equation, we can determine curvature scalar  $R$  in terms of  $\Lambda$ , and then derive the relation

$$R_{\mu\nu} = \frac{\Lambda}{2-D} g_{\mu\nu} \quad (7.13)$$

In this case the Ricci tensor is proportional to the metric and these are Einstein spaces. This is also maximally symmetric space [129] with the property that

$$R_{\mu\nu\rho\lambda} = \frac{R}{D(D-1)}(g_{\nu\lambda}g_{\mu\rho} - g_{\nu\rho}g_{\mu\lambda}) \quad (7.14)$$

The example of such space, with nonzero curvature, are de Sitter, Anti-de Sitter and D-spheres. In this sign convention, AdS space has *positive* cosmological constant. The AdS space is best described by an embedding. We start with  $D+1$  dimensional pseudo-Euclidean embedding space with coordinates  $\{y^a = y^0, y^1, \dots, y^{D-1}, y^D\}$  and metric  $\eta = \text{diag}(+, -, -, \dots, +)$  and the distance squared is

$$y^2 \equiv (y^0)^2 + (y^D)^2 - \sum_{n=1}^{D-1} (y^n)^2 \quad (7.15)$$

Note the appearance of two time coordinates from the form of the metric. The length remains invariant under  $SO(D-1, 2)$  global transformations

$$y^n \rightarrow y'^n = L_m^n y^m \quad (7.16)$$

where  $L_m^n$  is an  $SO(D-1, 2)$  matrix. If we consider the locus of

$$y^2 = b^2 = \text{constant} \quad (7.17)$$

and that defines  $AdS_D$ . It is worth noting that the invariance group for theories defined on  $AdS_D$  is same as that of the D-dimensional flat space that is D generators corresponding to translations and  $\frac{1}{2}D(D-1)$ , generators from Lorentz rotations.

Next, let us consider what is the conformal group in D-dimensional Euclidean space  $E^n$ . In this case the Poincare group has altogether  $\frac{1}{2}D(D+1)$  generators (D translations and rest from Lorentz group). Then we have following extra generators:

$$\vec{x} \rightarrow \lambda \vec{x} \quad (7.18)$$

this is *dilation* and  $\lambda$  is a real number. Furthermore, there is special conformal transformation

$$\frac{x'^\mu}{x'^2} = \frac{x^\mu}{x^2} + \alpha^\mu \quad (7.19)$$

This transformation involves n parameters  $\alpha^\mu$  The transformation (7.19) can be rewritten as

$$x'^\mu = \frac{x^\mu + \alpha^\mu x^2}{1 + 2x^\mu \alpha_\mu + \alpha^2 x^2} \quad (7.20)$$

Thus we see that the total number of generators are:  $\frac{1}{2}D(D+1) + 1 + D = \frac{1}{2}(D+1)(D+2)$ . This is the same number of generators that  $AdS_{D+1}$  space has. Indeed, in view of the recent developments, one can establish the connection that the isometry group of  $AdS_{D+1}$ ,  $SO(2, D)$  acts on the boundary as the conformal group acting on Minkowski/Euclidean space. We list below the generators of conformal group and their algebra

$$[M_{\mu\nu}, P_\lambda] = i(g_{\nu\lambda}P_\mu - g_{\mu\lambda}P_\nu) \quad (7.21)$$

$$[M_{\mu\nu}, M_{\lambda\rho}] = i(g_{\mu\rho}M_{\nu\lambda} - g_{\mu\lambda}M_{\nu\rho} + g_{\nu\lambda}M_{\mu\rho} - g_{\nu\rho}M_{\mu\lambda}), \quad (7.22)$$

$$[M_{\mu\nu}, K_\lambda] = i(g_{\nu\lambda}K_\mu - g_{\mu\lambda}K_\nu) \quad (7.23)$$

$$[D, P_\mu] = iP_\mu \quad (7.24)$$

$$[D, K_\mu] = -iK_\mu \quad (7.25)$$

$$[P_\mu, K_\nu] = 2i(g_{\mu\nu} + M_{\mu\nu}) \quad (7.26)$$

The generators of conformal transformation have the following representations, when we choose Cartesian coordinate system and consider transformation properties of a real scalar field:  $P_\mu = -i\partial_\mu$ ,  $M_{\mu\nu} = x_\nu P_\mu - x_\mu P_\nu$ ,  $D = x^\mu P_\mu$  and  $K_\mu = x^2 P_\mu - 2x_\mu D$ , corresponding to translation, Lorentz transformation, dilation and special conformal transformations respectively.

Let us discuss the evidences in support of Maldacena's conjecture. When we consider collections of D3-branes of the type IIB theory we note that D3-branes couple to the 5-form field strength and  $N$  units of this flux will pass through the five sphere of the  $AdS_5 \times S_5$  manifold. The isometry group of  $S_5$  is  $SO(6)$  and the  $AdS_5$  is endowed with isometry group  $SO(4,2)$  as we have just mentioned. The IIB theory has fermions and therefore, it is more relevant to consider the covering groups  $SU(4)$  and  $SU(2,2)$  of  $SO(6)$  and  $S(4,2)$  respectively. We also know that type IIB theory has 32 Majorana supercharges. These supersymmetries are preserved by the background under consideration. The invariance group is the super Lie group  $SU(2,2|4)$  for this theory. On the super Yang-Mills part, one has to examine how the above symmetry appears on the boundary theory. We have mentioned how the conformal group, for the case at hand, is to be identified as  $SO(4,2)$  or  $SU(2,2)$ . It is well known that  $N = 4$  super Yang-Mills theory is conformally invariant in four dimensions, since the theory has vanishing  $\beta$ -function [130], and thus the origin of the conformal group is well understood. Let us now focus our attention on the other symmetries present in type IIB theory. The ten dimensional super Yang-Mills has gauge bosons,  $A_\mu^a$ ,  $\mu = 0, 1, \dots, 9$ ,  $a$  being  $U(N)$  group index and thus there are 8 physical states corresponding to each gauge field. The superpartners are Majorana Weyl gauginos having matching numbers. The theory has 16 Majorana supercharges in  $D = 10$ . When we consider the 4-dimensional action, dimensionally reduced from ten dimensions [48, 49] physical degrees of freedom of each of the ten dimensional gauge field decomposes into 2 (corresponding to physical degrees of freedom of gauge field in  $D = 4$ ) and six scalars,  $\phi_i^a$ ,  $i = 1, 2, \dots, 6$ ,  $a$  is group index suppressed from now on. The number of, gauginos are given by the Weyl spinors,  $\lambda_\alpha^A$ ,  $A = 1, 2, 3, 4$ ,  $\alpha = 1, 2$ . One of these fermions, together with the gauge field can be grouped define a vector superfield. The rest of the three spinors can be grouped with the scalars (which appeared after dimensional reduction) to define 3 chiral superfields. The 16 supercharges can be grouped into 4 sets of complex Majorana charges  $Q_\alpha^A, \bar{Q}_\alpha^A$ ,  $A = 1, 2, 3, 4$  and  $\alpha = 1, 2$ . These two supercharges transform as  $\{4\}$  and  $\{\bar{4}\}$  of the R-symmetry group  $SU(4)$ . The scalars  $\phi_i$  transform as  $\{6\}$  of the  $SO(6)$ , since we deal with the covering group  $SU(4)$ , the scalars transform in the antisymmetric, rank 2 representation of the  $SU(4)$ . We see that type IIB theory has 32 supercharges, but the super Yang-Mills has only 16 of them. We know from discussions in Sec.IV that in the presence of the coincident D3-branes, half of the supersymmetries are preserved. When we consider the superconformal algebra the rest also appear as the extension of the superconformal group [131].

Another important nonperturbative symmetry of type IIB theory is the  $SL(2, Z)$  symmetry where dilaton and axion define the moduli. In the Yang-Mills sector the S-duality symmetry is robust and is known to be, again,  $SL(2, Z)$ . In this case, the modular parameter  $\tau = \frac{\theta}{2\pi} + \frac{4i\pi}{g_{YM}^2}$  whereas in the former case it is  $\tau = \chi + ie^{-\phi}$ .

The preceding discussions were focused to show that the symmetry properties of the type IIB theory and those of  $N = 4$  super Yang-Mills are the same. It is important to investigate which physical properties are common to both the theories. Indeed, if the two theories are equivalent, it should be possible to identify a physical field  $\Psi$  in the bulk theory and find the corresponding object on the boundary theory. Then, one of the tests will be to compute the correlators involving relevant objects in each of the theories and check the consistencies. Thus it is important to identify the physical quantities (operators) in both the theories. In the case of the boundary theory, one obvious criterion will be to choose gauge invariant operators while computing the correlators. One could formally express the equivalence between the theories through the relation among the generating functionals.

$$e^{-S_{II}[\Phi(J)]} = \int \mathcal{D}A e^{-(S_{YM}[A] + \mathcal{O}_\Delta[A]J)} \quad (7.27)$$

The l.h.s. of the above equation is to be identified as the generating function for the supergravity theory (rather low energy limit of IIB theory). The action  $S_{II}$  is determined in terms of the

massless states of the supergravity and the Kaluza-Klein towers and these are collectively denoted as  $\Phi(z, \omega)$ . Here the coordinates  $z^N \equiv (x^\mu, r)$  and  $\mu$  taking values 0,1,2,3 are to be identified as the AdS coordinates and  $\omega$  is the coordinate on five sphere. Moreover, it is implied due to the presence of  $J(x)$  that it also depends on the boundary data of the bulk fields. The *r.h.s.* defines the generating function for  $N = 4$  super Yang-Mills theory; however, one only computes the correlation functions of gauge invariant composite operators denoted by  $\mathcal{O}(A)$  with couplings to  $J(x)$ . In this general setting [122, 123, 132], one will be able to compute the correlation functions from both the theories and establish the correspondence between the two theories. Let us consider a simple example as illustration for the case of minimally coupled scalar in the bulk theory which could be identified with the dilaton. The action on the bulk for the dilaton on  $AdS_5 \times S_5$  is

$$S = \frac{\pi^3 b^3}{4G_{10}} \int d^5x \sqrt{|g|} g^{\mu\nu} \partial_\mu \phi \partial_\nu \phi \quad (7.28)$$

The factor  $\pi^3 b^3$  comes from the volume of  $S_5$ , through implicit assumption that  $\phi$  has no dependence on coordinates of five sphere. The metric is  $g_{\mu\nu} = \frac{b^2}{x_0^2} \delta_{\mu\nu}$ , is metric on  $AdS_5$ , now in the Poincare coordinates. For large  $\lambda \gg 1$ , the classical supergravity can be taken to be a good approximation (7.10). The dilaton equation of motion is given by

$$\partial_\mu (\sqrt{g} g^{\mu\nu} \partial_\nu \phi) = 0 \quad (7.29)$$

Of course, this equation can be solved by the standard Green's function method. The purpose is to determine the generating function with value of dilaton computed on the boundary, call it  $\phi_0$  which is value of  $\phi$  as  $x_0 \rightarrow 0$ . Thus we can write

$$\phi(x_0, \vec{x}) = \int d^4\vec{z} K(x_0, \vec{x}, \vec{z}) \phi_0(\vec{z}) \quad (7.30)$$

the vectors refer to four dimensional vectors on the boundary space and the Green's function is defined as,

$$K(x_0, \vec{x}, \vec{z}) \sim \frac{x_0^4}{[x_0^2 + (\vec{x} - \vec{z})^2]^4} \quad (7.31)$$

Now, one can insert the solution for  $\phi$  into the action to determine it at the classical value of dilaton

$$S = \frac{\pi^3 b^8}{4G_{10}} \int \frac{d^4\vec{x}}{x_0^3} \phi \partial_0 \phi|_\epsilon^\infty \quad (7.32)$$

$\epsilon$  is the cut off for the lower limit of integration. Once expression for  $\phi$  is inserted into the action, then it is possible to take cut off to zero and everything is finite. The action is given by

$$S \sim -\frac{\pi^3 b^8}{4G_{10}} \int d^4\vec{x} \int d^4\vec{z} \frac{\phi_0(\vec{x}) \phi_0(\vec{z})}{(\vec{x} - \vec{z})^8} \quad (7.33)$$

Then the generating function can be obtained by exponentiating this action. On the super Yang-Mills side, since it is a conformal field theory in four dimensions, the quadratic of Yang-Mills field strength has dimension 4 and product of two of the  $F^2$  terms behave as

$$\langle F^2(\vec{x}) F^2(\vec{z}) \rangle \sim \frac{N^2}{(\vec{x} - \vec{z})^8} \quad (7.34)$$

If we want to determine the dilaton correlation function on boundary, we compute

$$\frac{\delta^2 Z_{II}(\phi_0)}{\delta \phi_0(\vec{x}) \delta \phi_0(\vec{z})} \sim \frac{N^2}{(\vec{x} - \vec{z})^8} \quad (7.35)$$

Now comparing (7.34) and (7.35) we find that they are in agreement. If one considers, metric perturbation of the form  $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$  and then computes the two point correlation of this

perturbation on the brane taking the boundary limit; this correlation is identical to the correlation of stress energy momentum tensors (product of a pair of them; just as we took correlation of two  $F^2$  terms while identifying the dilaton two point functions).

Let us recall that the 't Hooft coupling  $\lambda \equiv Ng_{YM}^2$  and the length parameter  $b^4 = l_s^4 \lambda = 4\pi l_s^4 N g_{str}$  are related. If we hold  $\lambda$  fixed and let  $N \rightarrow \infty$ , then the string coupling tends to zero. Therefore, string perturbation theory can give reliable result in this limit. Thus, one can get a full quantum theoretic description of the Yang-Mills theory in the  $N \rightarrow \infty$  limit. Instead of holding  $\lambda$  fixed, if we allow it to take large values, then in the domain, where AdS radius is kept constant, the relevant limit is  $\alpha' \rightarrow 0$ . We know that in the zero slope limit the string theory goes over to supergravity theory. We saw the matching of AdS/CFT in this limit. But the consequences of Maldacena conjecture is very interesting in this regime, it tells us how the superconformal gauge theory in the  $N \rightarrow \infty$  limit behaves in strong coupling domain. Of course, the example we have been considering is the one where the  $\beta$ -function of the theory vanishes identically and therefore, it is not a realistic theory if we want to establish connection with supersymmetric gauge theories which have running coupling constants leading to asymptotic freedom. There are attempts to construct field theories which will have broken SUSY and conformal invariance (for example classical SQCD is scale invariant, but in the quantized theory scale invariance is broken). Witten [133] has proposed that one should consider  $AdS_7 \times S_4$ . The resulting boundary theory corresponds to 6-dimensional theory whose action is yet to be explicitly constructed. Then one compactifies the theory on  $T^2$  and require that fermions satisfy anti-periodic boundary condition around a cycle of the two-torus. Then the boundary theory is a four dimensional one. Conformal invariance and supersymmetry are broken in this 4-dimensional theory and we have a pure gauge theory with large  $N$ .

There has been rapid developments in studying the interconnection between supergravity (rather type IIB) theory on AdS space and boundary gauge theory. Several important issues pertaining to string theory and gauge theories have been addressed in this context. We refer to some of the recent review articles in this subject [134, 135, 136].

## 8 Summary and Conclusion

We have made some efforts to convey to the reader some of the interesting and important developments in string theory through this article. It is not possible to include all developments in the field in diverse directions in an article of this nature. A global perspective of string theory is contained in the article of John Schwarz [137] in this volume. We may recall that the research in string theory has stimulated progress in other fields such as mathematics, quantum field theory and statistical mechanics of lower dimensional systems to mention a few areas. We have seen that string theory has made very important contributions to our understanding of the physics of the black holes. As we have mentioned, for a special class of black holes, the Beckenstein-Hawking entropy formula could be derived from an underlying microscopic theory. Similarly, the nature of the Hawking radiation from a stringy black hole, slightly away from extremality, could be derived from the theory.

We have noted that, there are intimate connections between the five string theories. Some of them are inter related through dualities in ten dimensions and some are related in lower dimensions. Thus it is recognized that dualities have a special role in our understanding of string dynamics. Moreover, there are increasing evidence that there is a unique, fundamental theory and the five perturbatively consistent string theories are various phases of the fundamental theory. It is argued that M-theory might be that theory and the low energy effective action of M-theory is to be identified with the eleven dimensional supergravity theory. In this context, we discussed the M(atrrix) model proposal to show that the model captures many important features of M-theory.

Recently, the conjecture due to Maldacena has attracted considerable attention since it provides an important connection between supergravity on the bulk and the supersymmetric gauge theories living on the boundary. The connection between type IIB theory on  $AdS_5 \times S^5$  and  $N = 4$  supersymmetric gauge theory on the boundary has been at the center of attention. Furthermore, there are interesting developments in the study of theories on  $AdS_3$  and corresponding two dimensional

conformal field theories.

One of the most important achievements of string theories has been to address important issues in quantum gravity and provide answers to some of the puzzles. However, the theory is yet to provide a satisfactory answer to the cosmological constant problem. The cosmological constant is a parameter in physics which is measured to be closest to zero. It plays a dual role. When we look at it from the point of view of macroscopic physics, the smallness of the constant conveys to us that the Universe is very large and it is flat. On the other hand, it is expected that, the cosmological constant, like other parameters in Nature, should be explained from a microscopic theory and the short distance physics, i.e. quantum gravity, will explain the smallness of the cosmological constant. Therefore, one expects that string theory will be able to resolve this outstanding problem [138, 139]. The author along with his collaborators had made an attempt in this direction [140]. It is expected that string theory will provide us clues to understand the creation of the Universe and the evolution of the Universe in early epochs. Indeed, string cosmology has attracted considerable attention in recent years; however, we have not included discussions on this topic in this article due to limitations of space. Indeed, string cosmology makes several predictions which might be subjected to experimental tests in next few years [141].

### Acknowledgments

I would like to thank P. Majumdar, S. Panda, B. Sathiapalan and J. H. Schwarz for their suggestions and advice. I would like to thank the Yukawa Institute for Theoretical Physics, Professor Maskawa, Ninomiya and Sasaki for their very warm hospitality where most of this article was written.

### References

1. G. Veneziano, *Nuovo Cimento*, **57A**,190(1968)
2. E. Donini and S. Scuito, *Ann. Phys.* **58**, 388(1970).
3. D. B. Fairlie and H. B. Nielsen, *Nucl. Phys.* **B20**, 637(1970); C. S. Hsue, B. Sakita and M. A. Virasoro, *Phys. Rev.* **D2**, 2857(1970).
4. Y. Nambu, 'QuarkModel and Factorization of Veneziano Amplitude', in *Symmetries and quark model*, Ed. R. Chand (Gordon and Breach), 1970; H. B. Nielsen, 'An almost physical interpretation of the integrand of the n-point Veneziano amplitude', Submitted to the 15th International Conference on High Energy Physics, (Kiev); L. Susskind, *Nuovo Cim.* **69A**, 457(1970).
5. M. A. Virasoro, *Phys. Rev.* **177**, 177(1969).
6. J. Shapiro, *Phys. Lett.* **33B**, 361(1970).
7. J. Scherk and J. H. Schwarz, *Nucl. Phys.* **B81**, 118(1974).
8. M. B. Green and J. H. Schwarz, *Phys. Lett.* **B149**, 117(1984).
9. E. Witten, *Phys. Lett.* **B149**, 35(1984).
10. D. J. Gross, J. A. Harvey, E. Martinec and R. Rohm, *Phys. Rev. Lett* **54**, 502(1985); *Nucl. Phys.* **B256**,253(1985); *Nucl. Phys.* **B267**,75(1986).
11. M. B. Green, J. H. Schwarz and E. Witten, *Superstring Theory*, Vol I and Vol II, Cambridge University Press, 1987.

12. J. Polchinski, String Theory, Vol I and Vol II, Cambridge University Press, 1998.
13. V. Alessandrini, D. Amati, M. Le Bellac and D. I. Olive, Phys. Rep. **1C**, 269(1971).
14. J. H. Schwarz, Phys. Rep. **8C**, 269(1973).
15. G. Veneziano, Phys. Rep. **12C**, 1(1974).
16. C. Rebbi, Phys. Rep. **12C**, 259(1974).
17. J. Scherk, Rev. Mod. Phys. **47**, 123(1975).
18. J. H. Schwarz, Phys. Rep. **89C**, 223(1982).
19. Int. J. Mod. Phys.**A9**, 3707(1994).
20. A. Giveon, M. Porrati and E. Rabinovici, Phys. Rep. **C244**, 77(1994).
21. M. Duff, R. Khuri, and J. Lu, Phys. Rep. **259C**, 213(1995).
22. S. Chaudhury, C. Johnson and J. Polchinski, hep-th/9602052.
23. J. H. Schwarz, Nucl. Phys. Suppl. **B55**, 1(1997).
24. J. Polchinski, Rev. Mod. Phys. **68**, 1245(1996).
25. M. J. Duff, hep-th/9611203.
26. P. K. Townsend, hep-th/9612121.
27. M. Douglas, hep-th/9610041.
28. P. K. Townsend, gr-qc/9707012; hep-th/9712004.
29. C. Vafa, hep-th/9702201.
30. E. Kiritsis, hep-th/9708130.
31. D. Youm, hep-th/9710046.
32. T. Banks, hep-th/9710231.
33. D. Bigatti and L. Susskind, hep-th/9712072.
34. Y. Nambu, 'Duality and Hydrodynamics', Lecture at the Copenhagen Symposium, 1970; T. Goto, Prog. Th. Phys. **46**, 1560(1971); Y. Hara, Prog. Th. Phys. **46**, 1549(1971).
35. A. M. Polyakov, Phys. Lett. **103B**, 207(1981).
36. S. Deser and B. Zumino, Phys. Lett. **62B**, 369(1976); L. Brink, P. Di Vecchia and P. Howe, Phys. Lett. **65B**, 471(1976); A. M. Polyakov, Phys. Lett. **103B**, 211(1981).
37. S. Fubini, J. Maharana, M. Roncadelli and G. Veneziano, Nucl. Phys. **B316**, 36(1989).
38. A. Neveu and J. H. Schwarz, Phys. Rev. **D4**, 1109(1971), Nucl. Phys. **B31**, 86(1971).
39. P. Ramond, Phys. Rev. **D3**, 2415(1971).
40. F. Gliozzi, J. Scherk and D. Olive, Phys. Lett. **65**, 282(1976); Nucl. Phys. **B122**, 253(1977).
41. C. Bachas, Lectures on D-branes, hep-th/9806199.
42. A. Sen, Int. J. Mod. Phys. **A9**, 3707(1994).
43. S. Fubini and G. Veneziano, Nuovo Cimento, **67A**, 29(1970).

44. L. Alvarez Gaume, D. Z. Freedman and S. Mukhi, *Ann. Phys.* **134**, 85(1981).
45. M. B. Green, J. H. Schwarz, L. Brink, *Nucl. Phys.* **B198**, 474(1982); K. Kikkawa and M. Yamazaki, *Phys. Lett.* **149B**, 357(1984); N. Sakai and I. Senda, *Prog. Th. Phys.* **75**, 692(1984); V. P. Nair, A. Shapere, A. Strominger and F. Wilczek, *Nucl. Phys.* **B287**, 402(1987).
46. A. Shapere and F. Wilczek, *Nucl. Phys.* **B320**, 669(189); A. Giveon, E. Rabinovici and G. Veneziano, *Nucl. Phys.* **B322**, 167(1989); A. Giveon, N. Malkin, E. Rabinovici, *Phys. Lett.* **220B**, 551(1989)
47. K. S. Narain, *Phys. Lett.* **169B**, 41(1986)
48. J. Scherk and J. H. Schwarz, *Nucl. Phys.* **B153**, 61(1979).
49. J. Maharana and J. H. Schwarz, *Nucl. Phys.* **B390**, 3(1993).
50. S. F. Hassan and A. Sen, *Nucl. Phys.* **B375**, 103(1992).
51. G. Veneziano, *Phys. Lett.* **265B**, 287(1991).
52. K. Meissner and G. Veneziano, *Phys. Lett.* **267B**, 33(1991).
53. K. Meissner and G. Veneziano, *Mod. Phys. Lett.* **A6**, 3397(1991)
54. M. Gasperini, J. Maharana and G. Veneziano, *Phys. Lett.* **272B**, 277(1992); *Phys. Lett.* **296B**, 51(1993).
55. A. Sen, *Phys. Lett.* **271B**, 295(1991); *Phys. Lett.* **272B**, 34(1992); *Phys. Rev. Lett.* **69**, 1006(1992).
56. G. 't Hooft, *Nucl. Phys.* **B79**, 276(1974).
57. A. M. Polyakov, *JETP Lett.* **20**, 194(1974).
58. C. Montonen and D. Olive, *Phys. Lett.* **72B**, 117(1977).
59. E. Witten, *Phys. Lett.* **86B**, 283(1979).
60. E. Witten and D. Olive, *Phys. Lett.* **78B**, 97(1978).
61. A. Font, L. Ibanez, D. Lust and F. Quevedo, *Phys. Lett.* **B249**, 35(1990). A. Shapere, S. Trivedi and F. Wilczek, *Mod. Phys. Lett.* **A6**, 2677(1991).
62. S. J. Rey, *Phys. Rev.* **D43**, 526(1991).
63. J. H. Schwarz and A. Sen, *Phys. Lett.* **B312**, 105(1993); *Nucl. Phys.* **B411**, 35(1994).
64. J. H. Schwarz, Dilaton-Axion Symmetry, Talk at the International Workshop on String Theory, Quantum Gravity and Unification of Fundamental Interactions, Rome, September 1992; hep-th/9209125.
65. A. Sen, *Phys. Lett.* **329B**, 217(1994).
66. N. Seiberg and E. Witten, *Nucl. Phys.* **B426**, 19(1995).
67. A. Dabholkar, G. Gibbons, J. A. Harvey and F. Ruiz Ruiz, *Nucl. Phys.* **B340**, 33(1990); A. Dabholkar and J. A. Harvey, *Phys. Rev. Lett.* **63**, 478(1989).
68. G. T. Horowitz and A. Strominger, *Nucl. Phys.* **B360**, 197(1991).
69. J. Polchinski, *Phys. Rev. Lett.* **75**, 4724 1996.



70. M. Duff, Nucl. Phys. **B442**, 47(1995).
71. C. Hull and P. K. Townsend, Nucl. Phys. **B438**, 109(1995).
72. M. Duff, Nucl. Phys. **B442**, 47(1995); M. Duff and R. Khuri, Nucl. Phys. **411**, 473(1994).
73. E. Witten, Nucl. Phys. **B443**, 85(1995).
74. A. Sen, Nucl. Phys. **B450**, 103(1995).
75. J. A. Harvey and A. Strominger, Nucl. Phys. **449**, 535(1995)
76. M. Dine, P. Huet and N. Seiberg, Nucl. Phys. **B322**, 301(1989); J. Dai, R. G. Leigh and J. Polchinski, Mod. Phys. Lett. **A4**, 2073(1989).
77. F. Giani and M. Pernici, Phys. Rev. **D30**, 325(1984); I. Campbell and P. West, Nucl. Phys. **B243**, 112(1984); M. Huq and M. Namazie, Class. Quant. Grav. **2**, 293(1985).
78. P. K. Townsend, Phys. Lett. **350B**, 184(1995).
79. M. J. Duff, P. S. Howe, T. Inami and K. S. Stelle, Phys. Lett. **191B**, 70(1987); M. J. Duff and K. Stelle, Phys. Lett. **253B**, 113(1991).
80. J. H. Schwarz, Phys. Lett. **367B**, 97(1996).
81. J. H. Schwarz, Phys. Lett. **360B**, 13(1995). J. Polchinski and E. Witten, Nucl. Phys. B **460**, 525 (1996); J. Polchinski, Rev. Mod. Phys. **68**, 1245(1996); M. J. Duff, Int. J. Mod. Phys. A **11**, 5623 (1996).
82. C. Hull and P. Townsend, Nucl. Phys. B **438**, 109 (1995).
83. E. Witten, Nucl. Phys. B **443**, 85 (1995).
84. P. Horava and E. Witten, Nucl. Phys. **B460**, 506(1996); Nucl. Phys. **B475**, 94(1996).
85. K. Dasgupta and S. Mukhi, Nucl. Phys. **465**, 399(1996).
86. E. Witten, Nucl. Phys. **B463**, 383(1996).
87. A. Sen, Mod. Phys. Lett. **A11**, 1339(1996).
88. J. Bardeen, B. Carter and S. W. Hawking, Comm. Math. Phys. **31**, 161(1973).
89. J. Beckenstein, Lett. Nuov. Cimento **4**, 737(1972); Phys. Rev. **D7**, 2333(1973); Phys. Rev. **D9**, 3292(1974).
90. S. W. Hawking, Nature **248**, 30(1974); Commun. Math. Phys. **43**, 199(1975).
91. L. Susskind and J. Uglam, Phys. Rev. **D50**, 2700(1994); J. Russo and L. Susskind, Nucl. Phys. **B437**, 611(1997).
92. S. Fubini and G. Veneziano, Nuovo Cimento **64A**, 811(1970)
93. A. Sen, Mod. Phys. Lett. **A10**, 2081(1995).
94. A. Strominger and C. Vafa, Phys. Lett. **379B**, 99(1996).
95. E. Witten, Nucl. Phys. **B460**, 335(1996).
96. A. Sen, Phys. Rev. **D54**, 2964(1996); Phys. Rev. **D53**, 2874(1996).
97. C. Vafa, Nucl. Phys. **B463**, 415(1996).
98. C. Vafa, Nucl. Phys. **B463**, 435(1996).

99. S. R. Das and S. D. Mathur, hep-th/9601152.
100. C. G. Callan and J. M. Maldacena, Nucl. Phys. **B472**, 591(1996).
101. J. M. Maldacena, Black holes in string theory, hep-th/9607235; this Princeton University thesis has a comprehensive presentation of black hole entropy and Hawking radiation derived from string theory.
102. S. R. Das and S. D. Mathur, Nucl. Phys. **478**, 561(1996); Nucl. Phys. **B482**, 153(1996).
103. S. Gubser and I. Klebanov, Nucl. Phys. **B482**, 173(1996); Phys. Rev. Lett. **77**, 4491(1996).
104. A. Dhar, G. Mandal and S. R. Wadia, Phys. Lett. **388B**, 51(1996).
105. T. Banks, W. Fischler, S. H. Shenker and L. Susskind, Phys. Rev. **D55**, 112(1997)
106. A. Bilal, M(atric) theory: a pedagogical introduction, hep-th/ 9710136.
107. S. Fubini and G. Furlan, Physics, **1**, 229(1965); S. L. Adler, Phys. Rev. Lett. **14**, 1051(1965).
108. S. Weinberg, Phys. Rev. **150**, 1313(1966).
109. J. Kogut and L. Susskind, Phys. Rep. **8C**, 75(1973).
110. U. H. Danielsson, G. Ferrari and B. Sundborg, Int. J. Phys. **A11**, 5463(1996); D. Kabat and P. Pouliot, Phys. Rev. Lett. **77**, 1004(1996).
111. R. P. Feynman, Photon Hadron Collisions, Benjamin, 1973.
112. S. Sethi and M. Stern, Commun. Math. Phys. **194**, 675(1998)
113. M. Porrati and A. Rozenberg, Nucl. Phys. **B515**, 184(1998).
114. K. Becker and M. Becker, Nucl. Phys. **B506**, 48(1997).
115. A. Achucarro, J. M. Evans, P. K. Townsend and D. L. Wiltshire, Phys. Lett. **198B**, 441(1987);
116. B. de Wit, M. Lüscher and H. Nicolai, Nucl. Phys. **B305** [FS23], 545(1988).
117. T. Banks and N. Seiberg, Nucl. Phys. **B497**, 41(1997); R. Dijkgraaf, E. Verlinde and H. Verlinde, Nucl. Phys. **B500**, 43(1997).
118. N. Ishibashi, H. Kawai, Y. Kitazawa and A. Tsuchiya, Nucl. Phys. **B498**, 467(1997)
119. Y. Makeenko, Three Introductory Lectures in Helisinki on Matrix Models of Superstrings, hep-th/9704075.
120. J. Maldacena, Adv. Theor. Math. Phys. **2**, 231(1998).
121. G. 't Hooft, Nucl. Phys. **B72**, 461(1974).
122. S. S. Gubser, I. Klebanov and A. M. Polyakov, Phys. Lett. **428B**, 105(1998).
123. E. Witten, Adv. Theor. Math. Phys. **2**, 253(1998).
124. E. Witten and L. Susskind, The Holographic Bound in Anti-de Sitter Space, hep-th/9805114.
125. C. R. Stephens, G. 't Hooft and B. F. Whiting, Class. Quant. Grav. **11**, 621(1994); G. 't Hooft gr-qc/9310026.
126. L. Susskind, J. Math. Phys. **36**, 6377(1995).
127. A. M. Polyakov, Nucl. Phys. **B68**, 1 (1998); Proc. Suppl.

- 128. A. Tseytline, Nucl. Phys. **B501**,41 (1997).
- 129. S. Weinberg, Gravitation and Cosmology,
- 130. S. Mandelstam, Nucl. Phys. **B213**, 149(1983).
- 131. R. Haag, J. T. Lopuszanski and M. Sinius, Nucl. Phys. **B88**, 257(1975).
- 132. F. Ferrara, C. Fronsdal and A. Zaffaroni, Nucl. Phys. **B532**, 153(1998).
- 133. E. Witten, Adv. Theor. Math. Phys. **2**, 505(1998).
- 134. J. L. Petersen, Introduction to the Maldacena Conjecture on AdS/CFT, hep-th/9902131.
- 135. P. Di Vecchia, An Introduction to AdS/CFT equivalence, hep-th/9903007.
- 136. M. B. Green, Interconnections between type II superstrings, M theory and  $\mathcal{N} = 4$  supersymmetric Yang-Mills, hep-th/9903124.
- 137. J. H. Schwarz, in this Volume.
- 138. E. Witten, Mod. Phys. Lett. **A10**, 2153(1995).
- 139. K. Becker, M. Becker and A. Strominger, Phys. Rev. **D51**, 6603(1995).
- 140. S. Kar, J. Maharana and H. Singh, Phys. Lett. **B374**, 43(1996).
- 141. G. Veneziano, CERN Preprint, CERN-TH/98-43, hep-th/9802057.

# 25. Yang–Mills Theory and Matrix String Theory

L. Bonora \*

International School for Advanced Studies (SISSA/ISAS)  
Via Beirut 2–4, 34014 Trieste, Italy, and INFN, Sezione di Trieste

## Abstract

This is a review of some recent developments in the study of the relation between Yang–Mills theory and strings. The relation we are concerned with here is based on classical Yang–Mills theory solutions called *Riemannian instantons*. The latter are two-dimensional solutions describing, in the strong coupling, Riemann surfaces. They lend therefore themselves to an interpretation in terms of string theory interaction. This interpretation is worked out in detail for the so-called Matrix String Theory, i.e. for the 2d Yang–Mills theory with  $\mathcal{N} = (8, 8)$  supersymmetry obtained as reduction of the  $\mathcal{N} = 1$  Yang–Mills theory with gauge group  $U(N)$  in 10d. In fact it is argued that, in the strong coupling limit, the Matrix String Theory describes type IIA superstring theory.

## 1 Introduction

The rich structure of non-Abelian gauge theories is at the core of their successful employment as theories of the elementary particles. It is this complicated structure that allows us to claim to be able, at least potentially, to describe in a consistent way by the same theory both confinement and asymptotic freedom, quark, gluons, mesons and hadrons. Classical and quantum non-Abelian gauge theories have been analyzed in countless papers in the past, but they do not cease to surprise us by revealing from time to time new structure. This has happened also recently in connection with the so-called second string revolution. Gauge theories have been naturally associated with branes and this association has revealed new, previously unsuspected features. This review is devoted to one such new development: the connection between Yang–Mills theories and strings via classical configurations called Riemannian (*stringy*) instantons.

Actually the connection between Yang–Mills theories and strings is far from new and seems to be multiform. That a link should exist was recognized long ago by 't Hooft, [1]. In this case Riemann surfaces appeared as an auxiliary structure underlying the relevant Feynman diagrams and led to the  $1/N$  expansion. This gave rise to a vast literature, see [3], especially in 2d, where the connection with strings became more concrete: see for example  $QCD_2$ , analyzed for its string-like properties [2]. More recently another connection was found between conformal SYM and IIB supergravity/superstring theory at large  $N$  in the AdS geometrical framework, [4]. Of a similar nature is the link between non-supersymmetric Yang–Mills theories and type 0 strings, [6], as well as the non-critical string approach, pioneered by Polyakov, [5]. All these are duality relations, i.e. they relate a Yang–Mills theory in a given range of the coupling to a string theory in a related range of the string coupling.

In this review we deal with a more direct link between string theory and non-Abelian Yang–Mills theory: not a duality relation between strings and Yang–Mills theories, but the emergence in the latter of classical solutions modeled over Riemann surfaces, which naturally lend themselves to a string interpretation. Such configurations exist in any non-Abelian Yang–Mills theory in dimensions 4 or higher, and even in 2d when adjoint matter is present. As will be seen, in some instances the string interpretation is clear. In other instances further analysis is needed.

---

\*Email: bonora@sissa

This string-like nature of Yang–Mills theories is something that could have been found long ago, but in fact the first example was brought to light only after the proposal of Matrix Theory [9]. The latter, in the large  $N$  limit, is expected to describe M theory. Therefore upon compactifying it on a circle one should end up, in the appropriate limit, with type IIA superstring theory [10, 11, 12, 13]

Now, by compactifying Matrix Theory on a circle we obtain (see below)  $\mathcal{N} = (8, 8)$  super–Yang–Mills (SYM) on a cylindrical 2D space-time with gauge group  $U(N)$ . Therefore the conjecture, supported with various arguments [10, 11, 13], is that this theory represents in the strong Yang–Mills coupling limit a theory of type II superstrings (see also [14, 15, 16]). Hereafter we refer to this theory as Matrix String Theory (MST).

Let us briefly review how decisive evidence in favor of this identification has been found recently. A first step in this direction was made in refs. [17, 18, 19], where it was pointed out that MST contains BPS instanton solutions which interpolate between different initial and final string configurations via suitable punctured Riemann surfaces. We often refer to them as *stringy* or *Riemannian instantons*. Subsequently, [20], it was shown that, in the strong coupling limit, MST in the background of a given classical BPS instanton solution reduces to the Green–Schwarz superstring theory plus a decoupled Maxwell theory, and that the leading term of the amplitude in such background is proportional to  $g_s^{-\chi}$ , where  $g_s$  is the string coupling constant (i.e. the inverse of the Yang–Mills coupling) and  $\chi = 2 - 2h - n$  is the Euler characteristic of the Riemann surface of genus  $h$  with  $n$  punctures, which characterizes the given classical solution. This is the result one expects from perturbative string interaction theory. Needless to say this is a strong confirmation of the abovementioned conjecture. Similar results have been seen to hold, [22], also for the Heterotic Matrix String Theory (HMST), i.e. for  $\mathcal{N} = (8, 0)$  SYM in 2d with gauge group  $O(N)$ : in the strong Yang–Mills coupling limit one finds the heterotic superstring theory with suitably broken gauge group.

This is not the end of the story, as far as the identification of strong coupling SYM and superstring theories is concerned. One cannot limit oneself to a qualitative correspondence between the two types of theories. In the path integral we have to integrate over all possible Riemannian instantons (modulo symmetries), therefore the question of instanton moduli is of paramount importance. In fact a few precise things can be said in this regard. In [21], the study of the moduli space of Riemannian instantons was taken up. Any such instanton consists of two ingredients, a *group theoretical factor* and a *core*. The latter corresponds to a branched covering of the cylinder. The group theoretical factor contains fields that satisfy WZNW-like equations with delta-function sources. Inside these instantons, Riemann surfaces appear as branched coverings of the base cylinder in the form of *plane curves*, i.e. the zero locus of order  $N$  polynomials of two complex variables. One can then set out to study the moduli space of such curves. This problem is of utmost importance because it turns out to be connected with the discrete parameter  $N$  and, therefore, it affects the large  $N$  limit. As we will see, for finite  $N$ , the instantons of MST reproduce exactly only tree string amplitudes, while they cover only part of the moduli space of higher genus Riemann surfaces with punctures. More precisely, in a process with  $n$  external string states mediated by a Riemann surface of genus  $h$ , one expects  $3h - 3 + n$  complex moduli; at finite  $N$ , in MST,  $h$  of them are discrete. In [21] it was argued that, when  $N \rightarrow \infty$ , these discrete moduli become continuous and one can recover the full moduli space of Riemann surfaces only.

After this historical reconstruction a clarification is in order. A thorough discussion of the relation between Yang–Mills theories and strings will be carried out here for SYM theories in 2d with a large amount of supersymmetry. However Riemannian instantons exist also with less or no supersymmetry at all in Yang–Mills theories in 4 or higher dimensions. Although in these cases the path integral impact of such configurations is not yet clear and, anyway, more complicated to evaluate, they represent an intriguing and not yet studied aspect of Yang–Mills theories. To stress this point we shall start in section 2 with the example of Riemannian instantons in 4d Yang–Mills theory.

The purpose of this review is to provide a pedagogical introduction and synoptic view of the above problems and results, which otherwise can be found scattered in various papers. The paper is organized as follows. In the next section we discuss a simple but significant example of Riemannian

instanton in Yang-Mills theory. The next section is a very sketchy introduction to type IIA theory, M theory and Matrix theory. Section 4 contains a formulation of MST. In section 5 it is explained how to construct the most general Riemannian instanton in MST. In section 6 it is shown that MST in the background of a given classical BPS instanton, characterized by a given Riemann surface, reduces to the Green-Schwarz superstring theory plus a decoupled Maxwell theory, and that the leading term of the amplitude in such background is proportional to  $g_s^{-\chi}$ ,  $\chi$  being the Euler characteristic of the Riemann surface in question. Section 7 is devoted to plane curves with a particular attention to singularities and their meaning. Finally in section 8 we concentrate on the moduli space of stringy instantons and describe the above-mentioned discretization. Two Appendices are devoted to the most technical aspects of this review.

## 2 Riemannian instantons in Yang-Mills theories

Let us consider, as an example, a pure Euclidean Yang-Mills theory in 4d with gauge group  $U(N)$ . Let  $A$  be the gauge connection form with curvature  $F$ . We use the hermitean convention for the Lie algebra-valued matrices so that the covariant derivative is  $D = d + igA$ , where  $g$  is the Yang-Mills coupling. We concentrate on the self-duality condition for the Yang-Mills field strength

$$F_{\mu\nu} = \frac{1}{2}\epsilon_{\mu\nu\lambda\rho}F^{\lambda\rho}$$

Rewritten in complex coordinates  $w = x^1 + ix^2$ ,  $\bar{w} = x^1 - ix^2$ ,  $y = x^3 + ix^4$ ,  $\bar{y} = x^3 - ix^4$ , it becomes

$$F_{w\bar{w}} = F_{y\bar{y}}, \quad F_{w\bar{y}} = 0 = F_{\bar{w}y} \quad (1)$$

To conform with the convention of MST, which will be used in the following sections, we suppose that the coordinate  $x^1, x^2$  span an infinite cylinder (precisely  $x^2$  is the periodic coordinate). We want to single out solutions of these equations which are independent of  $y, \bar{y}$ . Then, introducing the notation  $X = A_y, \bar{X} = A_{\bar{y}} = X^\dagger$ , (1) becomes

$$\begin{aligned} F_{w\bar{w}} - ig^2[X, \bar{X}] &= 0 \\ D_w \bar{X} &= 0, \quad D_{\bar{w}} X = 0 \end{aligned} \quad (2)$$

We refer to these equations as *Riemannian instanton equations* or, simply, *instanton equations*.

As an introductory example let us consider the case in which the gauge group is  $U(2)$  ( $N = 2$ ). We look for a couple  $(A, X)$  that satisfies eqs.(2). To this end we choose the following ansatz

$$X = Y^{-1}MY, \quad A_w = i\partial_w Y^\dagger(Y^{-1})^\dagger, \quad (3)$$

where  $Y$  is a suitable matrix  $\in SL(2, \mathbb{C})$ , and  $M$  is the following  $2 \times 2$  matrix

$$M = \begin{pmatrix} 0 & a \\ 1 & 0 \end{pmatrix} \quad (4)$$

where  $a$  is a function of the point in the  $w$  cylinder. As a consequence of the equation  $D_{\bar{w}}X = 0$ , it follows that  $\partial_{\bar{w}}a = 0$ , i.e.  $a$  is holomorphic in  $w$  (except perhaps at infinity on the cylinder). Now, given such a holomorphic  $a$  we want to find  $Y$  so that (2) is satisfied. We parametrize  $Y$  as follows:

$$Y = \begin{pmatrix} e^p & 0 \\ 0 & e^{-p} \end{pmatrix} = KL, \quad L = \begin{pmatrix} e^{\frac{u}{2}} & 0 \\ 0 & e^{-\frac{u}{2}} \end{pmatrix}, \quad K = \begin{pmatrix} |a|^{\frac{1}{4}} & 0 \\ 0 & |a|^{-\frac{1}{4}} \end{pmatrix} \quad (5)$$

where  $u$  is a function to be determined and  $p = \frac{u}{2} + \frac{1}{4}\ln|a|$ . Then using (3) we find

$$X = \begin{pmatrix} 0 & ae^{-2p} \\ e^{2p} & 0 \end{pmatrix}, \quad A_w = i\partial_w p \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad (6)$$

Now it is easy to verify that the first equation in (2) implies

$$2\partial_w\partial_{\bar{w}}p - g^2(e^{4p} - |a|^2e^{-4p}) = 0 \quad (7)$$

Inserting the explicit form of  $p$  and the change of variable  $w \rightarrow \zeta$ , s.t.  $\frac{\partial \zeta}{\partial w} = \sqrt{a}$ , one can rewrite (7) as

$$\partial_\zeta\partial_{\bar{\zeta}}u - 2g^2\sinh u = -\frac{\pi}{4}\delta(a)(\partial_\zeta a)(\partial_{\bar{\zeta}}\bar{a}). \quad (8)$$

where the derivatives are understood in the sense of complex distribution theory. This is the sinh-Gordon equation with delta-function-like boundary conditions at the points where  $a$  vanishes. If  $u$  is a smooth solution of this equation, the couple  $(X, A)$  is a solution of (2) which is smooth everywhere except perhaps at infinity on the cylinder. As we will see later such solutions exist, therefore solutions of the instanton equations do exist. However before we go to more details on the existence of solutions, it is important to discuss their meaning.

There are two distinct ingredients in (3): one is the matrix  $M$  (the *core*) and the other is the group theoretical element  $Y$ . We discuss them in turn.

## 2.1 Branched coverings

The matrix  $M$  represents a branched covering of the cylinder spanned by the coordinate  $w$ . In order to see this we diagonalize it by means of a matrix in  $SL(2, \mathbb{C})$ :

$$M = S\hat{M}S^{-1}, \quad \hat{M} = \begin{pmatrix} \sqrt{a} & 0 \\ 0 & -\sqrt{a} \end{pmatrix}, \quad S = \frac{i}{\sqrt{2}} \begin{pmatrix} a^{\frac{1}{4}} & a^{\frac{1}{4}} \\ a^{-\frac{1}{4}} & -a^{-\frac{1}{4}} \end{pmatrix}. \quad (9)$$

It is convenient now to pass to a new coordinate  $z = e^w$ , which maps the cylinder into the complex  $z$ -plane with two punctures at  $z = 0$  and  $z = \infty$ . The two eigenvalues of  $M$  (or, equivalently, of  $X$ ), which are the two roots of the algebraic equation  $X^2 = a$ , can be thought of as the sheets of a double covering of the cylinder. Each sheet is a copy of the complex  $z$ -plane. Each point on each sheet project to the corresponding point on the  $z$ -plane. Such projection will be denoted  $\pi$ . Suppose, for simplicity, that  $a = z - z_0$ . Then we have a branch point at  $z = z_0$  and another at  $z = \infty$ . We can draw a cut between these two branch points (for definiteness the cut will lie in the region  $|z| > |z_0|$ ). The two sheets are connected through the cut. What we mean by this, as is well known, is that if we consider going around the origin in the  $z$ -plane along a small circle of radius  $< |z_0|$ , we produce an inverse image (under  $\pi$ ) on the covering formed by two small circles around the origin, one for each sheet. But if we do the same operation for a circle with radius  $> |z_0|$ , we are bound to cross the cut: crossing the cut multiplies the root by a phase  $e^{i\pi}$ , so that  $\sqrt{a}$  goes over to  $-\sqrt{a}$  and viceversa, i.e. by crossing the cut we pass from one sheet to the other. This means that the counterimage of the circle on the covering this time is a long circle that extends over both sheets.

This result can be interpreted in two ways: a geometrical and a string theoretical one (although the latter is a bit premature at this stage). Geometrically what we have just described is a Riemann surface represented by a branched covering of the complex  $z$  plane. The Riemann surface in question has genus 0 and three punctures (see fig. 1, in all the figures of this paper Riemann surfaces are represented with finite boundaries in order to better suggest the string interpretation; however these boundaries are all asymptotic, they correspond to  $z = 0$  or  $z = \infty$ , so it is more appropriate to think of them as punctures).

The string interpretation is rather obvious. The coordinate  $x_1$  is taken to be the Euclidean time. Therefore  $z = 0$  correspond to time  $-\infty$  and  $z = \infty$  to time  $+\infty$ . It is natural to interpret the counterimages of the small circle around the origin as two incoming strings. This configuration does not change if we enlarge the circle around  $z = 0$  as long as its radius remains smaller than  $|z_0|$ . We say that the two strings propagate in time without interacting. If the radius of the circle in the

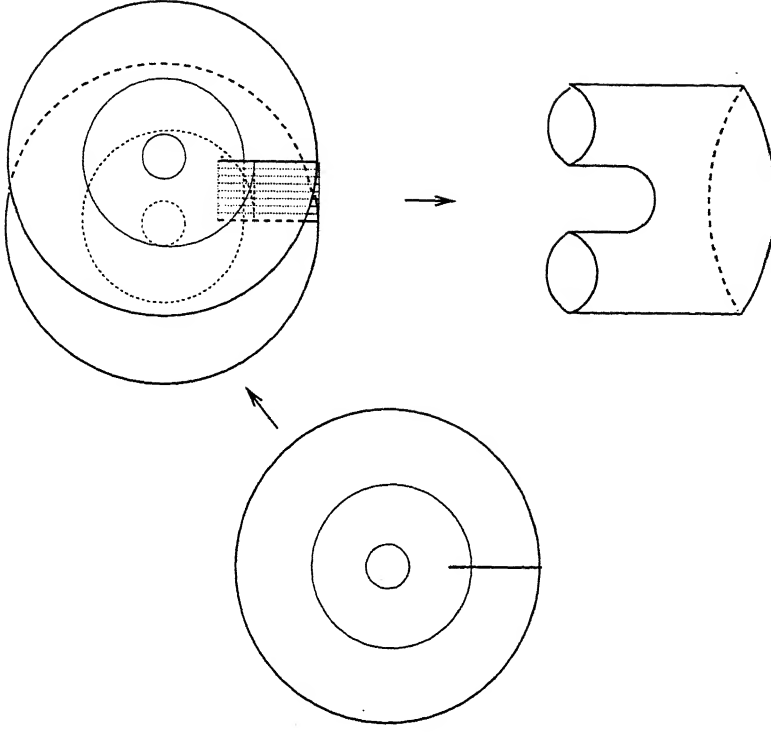


Figure 1: Simple branch

$z$  plane becomes larger than  $|z_0|$ , the counterimage becomes a unique closed curve extending over both sheets, and this configuration propagates unchanged as far as  $\infty$ . Therefore the instanton we are considering represents the joining of two strings to form one long string of double length. The joining interaction takes place at  $z = z_0$ . The inverse image under  $\pi$  of  $z = 0$  and  $z = \infty$  correspond to the points of the Riemann surface where the incoming strings enter and the outgoing string exits, respectively.

If there are more branch points beside the one considered above, it is not difficult to see how they may give rise to more complicated Riemann surfaces with handles (and more complicated string interactions). We will return later on to the problem of constructing more complicated Riemann surfaces and justifying the string interpretation outlined above. For the time being we would like to complete the description of our simple example by digging out from it any useful information. We notice that the passage through the cut can be described mathematically as the monodromy transformation

$$\hat{M} \rightarrow \Lambda \hat{M} \Lambda^{-1} = \begin{pmatrix} -\sqrt{a} & 0 \\ 0 & \sqrt{a} \end{pmatrix}, \quad \Lambda = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad (10)$$

$M$  is monodromy invariant since simultaneously  $S \rightarrow S\Lambda^{-1}$ .

Another useful remark is the following:  $X$  can be rewritten as follows

$$X = Y^{-1}MY = L^{-1}U\hat{M}U^{-1}L \quad (11)$$

where

$$U = \frac{i}{\sqrt{2}} \begin{pmatrix} \left(\frac{a}{a}\right)^{\frac{1}{8}} & \left(\frac{a}{a}\right)^{\frac{1}{8}} \\ \left(\frac{a}{a}\right)^{\frac{1}{8}} & -\left(\frac{a}{a}\right)^{\frac{1}{8}} \end{pmatrix} \quad (12)$$

It is a remarkable (and intentionally looked for) fact that  $U$  is a unitary matrix.  $U$  has the same monodromy as  $S$ :  $U \rightarrow U\Lambda^{-1}$ .



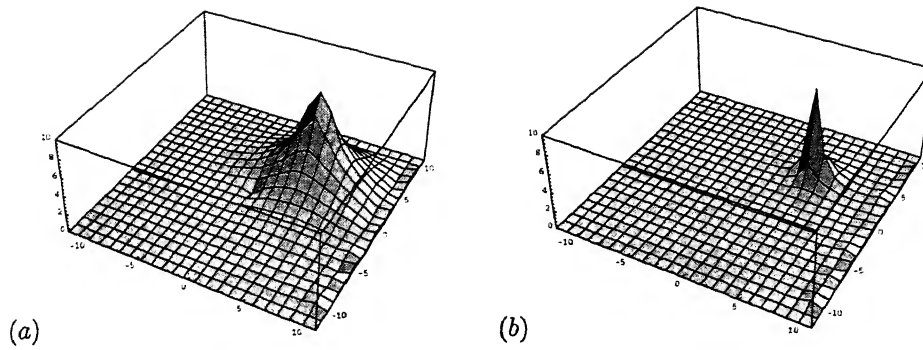


Figure 2: Numerical solutions of eq. (8) for (a) small and (b) large  $g$ , respectively.

## 2.2 The group theoretical factor

Let us now pass to the analysis of the group theoretical factor  $Y = KL$ . We have just seen that  $K$  concurs with the diagonalizing matrix  $S$  to form a unitary matrix  $U$ . We notice however that  $K$  diverges at  $z = z_0$ . Let us analyze next the *dressing factor*  $L$ . This amounts to returning to the solutions of eq.(8). The problem is to find a solution  $u$  that satisfies the sinh-Gordon equation with the boundary condition

$$u \sim -\frac{1}{2} \ln |a|, \quad a \sim 0. \quad (13)$$

It is thanks to this logarithmic singularity that the factors  $e^{\pm \frac{u}{2}}$  and  $K$ , contained in  $Y$ , compensate for the singularities of each other, so that the resulting solution  $(X, A)$  is smooth. Let us suppose again, for simplicity, that  $a = z - z_0$ . We want to single out, for this simple case, solutions of (8) with the right asymptotic behavior, that is behaving like (13) at  $z = z_0$  and vanishing at  $z = 0, \infty$ . First we rescale  $\zeta \rightarrow \sqrt{2}g\zeta$ , so that the sinh-Gordon equation takes the standard form

$$\partial_\zeta \partial_{\bar{\zeta}} u - \sinh u = 0. \quad (14)$$

We do not know an exact solution of this equation satisfying the boundary condition (13) and vanishing at the origin and at infinity of the  $z$ -plane. We can therefore proceed in two ways: find a numerical solution or an approximate analytical one. The latter, being rather technical, is discussed in Appendix A. The conclusion there is that the solution  $u$ , in the strong coupling limit, shrinks around the branch point  $z = z_0$ , that is it is practically zero everywhere except in a small neighborhood of this point where it has a spike-like behaviour.

This is confirmed by the numerical solutions shown in fig. 2.

We can easily extend the previous analysis to the case in which  $a$  contains several distinct zeroes. In the following we will simply assume that this is always the case, namely that in the strong coupling limit the dressing factor  $L$  tends very rapidly to 1 outside the branch points of the covering surface.

## 2.3 Summary

The purpose of this section was twofold: on the one hand, to show that in 4d Yang-Mills theory we have *Riemannian instantons*; on the other hand to work out an explicit simple example of such instantons. As for the first point it will become clear that Riemannian instantons, characterized by eqs.(2), exist in any Yang-Mills theory in dimensions  $\geq 4$  and also in 2d provided there is adjoint matter in the theory. The explicit solutions for the instantons is the same in all cases, but their role in the path integral may differ from case to case. From the example we have worked out above we learn how to proceed in order to find the general solution. A few things should be retained because they will turn out to hold in general: *a Riemannian instanton is made of two*

ingredients, the core constituted by a matrix  $M$  which defines a branched covering representing a Riemann surface with punctures, and a group theoretical factor. The latter splits into two factors: a dressing factor which tends to 1 in the strong coupling limit and the remnant, which concurs to form a unitary matrix when applied to  $M$ .

Our next step is to see the astonishing consequences of this fact at work in MST. Before going into that however, we have to prepare the ground with a short summary of purely pedagogical character.

### 3 IIA Superstring Theory and Matrix Theory

This section is a short review of type IIA superstring theory, M theory and Matrix theory, whose main purpose is to facilitate the comprehension of MST by inserting it in the historical framework in which it was first formulated.

Type IIA superstring theory can be formulated either in the Neveu–Schwarz–Ramond or in the Green–Schwarz (GS) formalism. For comparison with MST, the latter, in the light-cone gauge [7, 8], is the relevant one. In the light-cone gauge, out of ten fields  $x^M(\sigma, \tau)$  ( $M = 0, \dots, 9$ ), two are eliminated: defining  $x^\pm = x^0 \pm x^9$ , one sets  $x^+ \sim \tau$  to completely fix the gauge, while  $x^-$  turns out to be completely fixed in terms of the transverse degrees of freedom. So the true degrees of freedom of the theory are the transverse ones,  $x^i$  with  $i = 1, \dots, 8$ . To make a supersymmetric theory we have to add the fermionic part, which consists of two sets of 2D Majorana–Weyl fermions:  $\vartheta_s$  and  $\vartheta_c$ . They transform according to the  $8_s$  and  $8_c$  representation of  $SO(8)$ , respectively. The GS light-cone action is

$$S = -\frac{1}{2\pi} \int d\sigma d\tau \left( \partial_\mu x^i \partial^\mu x^i - i\bar{\vartheta} \rho^\mu \partial_\mu \vartheta \right), \quad (15)$$

where the integration range is  $-\infty < \tau < \infty$ ,  $0 < \sigma < \pi$ . In (15)  $\mu, \nu = 0, 1$ , and the 2D flat Minkowski metric  $\eta_{\mu\nu}$  is taken to have signature  $(-, +)$ .

Summation over the  $i$  index is understood. Moreover we have assembled  $\vartheta_s$  and  $\vartheta_c$  in a doublet  $\vartheta$  so that  $\vartheta^T = (\vartheta_s, \vartheta_c)$  and  $\bar{\vartheta} = \vartheta^T \rho^0$ . Finally  $\rho^\mu$  are the 2D gamma matrices.

The GS action (15) is invariant under two sets of rigid supersymmetry transformations, each characterized by 8 parameters. The massless spectrum of the GS superstring is the same as the spectrum of IIA supergravity in 10D, which constitutes its low energy effective theory. The bosonic massless fields include the dilaton  $\phi$ , the graviton  $g_{MN}$  and the antisymmetric tensor fields  $B_{MN}$  (the NSNS sector) together with the one-form  $A_M$  and the three-form  $C_{LMN}$  fields (the RR sector). Type IIA theory contains other brane-like objects, beside strings. They appear as macroscopic (soliton-like) solutions of IIA supergravity and their world-sheet couples to one of the potentials in the IIA supergravity theory. Therefore we have D0-branes and D2-branes, which couple to  $A$  and  $C$ , respectively, and D6-branes and D4-branes, which couple to their dual potentials. There is also a NS five-brane which couple to the potential dual to  $B$ . Our attention in the following will concentrate on fundamental strings and D0-branes.

Strings are the fundamental objects of type IIA theory, while, according to Matrix Theory, D0-branes are the fundamental constituents of M-theory. M-theory is closely related to IIA theory as follows. The type IIA string coupling  $g_s$ , i.e. the exponential of the vacuum expectation value of the dilaton, can be used as an expansion parameter when  $g_s$  is small – this is the region we will be concerned with in the following sections. When  $g_s$  becomes large the representation of IIA theory in terms of strings becomes unreliable. In fact as  $g_s \rightarrow \infty$  a radical change takes place in the theory: a new dimension opens up. This dimension is wound on a circle whose radius  $R$  in suitable units is given by  $g_s$ ; when  $g_s$  is small the additional dimension is dormant, when  $g_s$  is large it cannot be disregarded anymore. The new theory one lands on is eleven dimensional. This cannot be a string theory, because superstrings do not consistently propagate in 11D. It is an entirely new theory, M theory.

The low energy effective field theory of M theory is eleven dimensional supergravity. Its modes are the metric  $G$ , a 3-form tensor  $A$  and the gravitino, altogether 128 bosonic and 128 fermionic

degrees of freedom. M theory contains two types of soliton-like solutions, membranes and 5-branes. It is easy to see how the circle compactification maps the degrees of freedom of M theory into those of type IIA theory. For example, 5-branes of M theory generate NS 5-branes and D4-branes of IIA, membranes in M theory originate fundamental strings and D2-branes of IIA. In parallel, we can identify  $G_{11,11}$ ,  $G_{MN}$ ,  $A_{11MN}$  and  $A_{LMN}$  with  $\phi$ ,  $g_{MN}$ ,  $B_{MN}$  and  $C_{LMN}$ , respectively. Most important for us is however the origin of D0-branes and the potential  $A_M$  coupled to them. The latter is identified with  $G_{11M}$  and the D0-branes with the massive Kaluza-Klein (KK) modes of mass  $M = 1/g_s \sim 1/R$ . Actually the KK modes of 11D supergravity have mass  $n/g_s$  for any integer  $n$ . The case  $n = 0$  is not a D0-brane, but just the direct massless descendants of 11D supergravity. When  $n > 1$  the KK modes are interpreted as bound states at threshold of the elementary KK modes with  $n = 1$ . The integer  $n$  is identified with the D0-brane charge; negative  $n$ 's label anti-D0-branes. Each such massive KK mode comes in supermultiplet of 256 physical components.

For M theory there is not yet a handy formulation, similar, for example, to the one IIA theory has in terms of strings. At present there is a very radical proposal, still at an early stage of development: Matrix Theory (see [24, 25, 26, 27]). The basic objects of Matrix Theory are supergravitons, i.e. supermultiplets with 256 degrees of freedom to be identified with the D0-branes of IIA theory. To understand the Matrix Theory proposal we need to introduce preliminarily two new issues.

The first is the effective action for systems of  $N$  D-branes. D-branes are classically submanifolds of the target space where open strings can end (and therefore satisfy corresponding Dirichlet boundary conditions). The self-interaction and the mutual interactions of a system of D-branes is therefore representable in terms of string theory. The request of conformal invariance provides equations of motion for the D-brane fields that can be integrated and lead to a Born-Infeld type action. When the ambient metric is flat and the D-branes are almost flat (for D0-branes this means that their velocities are small compared to the velocity of light, which implies a non-relativistic framework) this action is well approximated by a super-Yang-Mills (SYM) action, obtained by simply reducing 10D SYM to the dimensions of the D-brane world-volume. In this non relativistic approximation the effective action appropriate for a system of  $N$  D0-branes is therefore 10D SYM reduced to one (time) dimension:

$$S_{D0} = \int dt \text{Tr} \left( \frac{1}{2g_s} (D_0 X^p)^2 - i\theta^T D_0 \theta + \frac{1}{4g_s} [X^p, X^q]^2 + \theta^T \hat{\Gamma}^p [X_p, \theta] \right) \quad (16)$$

where  $p = 1, \dots, 9$ ,  $X^p$  are  $N \times N$  Hermitean matrices,  $\hat{\Gamma}_p$  are  $16 \times 16$  gamma matrices of  $SO(9)$ ,  $\theta$  are  $N \times N$  matrices whose entries are 1D spinors. Moreover  $D_0 = \partial_0 - [A_0, \ ]$ . The quartic potential implies that, when D0-branes are far away from one another, the dominating configurations are given by diagonal  $X^p$  and  $\theta$ , the eigenvalues of  $X^p$  representing the distances among the various branes. When the D0-branes are nearby the non-diagonal terms become relevant and we are faced with a non-commutative space structure.

The second issue is the infinite momentum frame (IMF). Imagine a system of point particles with momenta  $p^a$ , where  $a$  labels the particles, and suppose we boost this system to very high velocity in a given direction. A simple calculation shows that the energy of the system takes the approximate form

$$E \approx p_{||}^{tot} + \sum_a \frac{(p_{\perp}^a)^2}{2p_{||}^a} \quad (17)$$

where the label  $||$  and  $\perp$  denote the momentum components parallel and perpendicular to the boost direction, respectively. Eq.(17) has a non-relativistic form.

Now let us consider a system of D0-branes and, with reference to (17), identify the transverse directions of this system with the nine space directions discussed above, while the boost direction is identified with the positive 11th dimension of M theory (so  $p_{||} = p_{11}$  and so on). When this system is boosted in the 11th direction we reach an effective non-relativistic regime. The idea of [9] is to assume that the nonrelativistic action (16) is an adequate description for such a system of  $N$

D0-branes boosted to infinite momentum. This would seem at first to be a proposal with limited scope. In fact an infinite boost in the positive 11th direction would seem to exclude from the game the D0-branes with  $n \leq 0$ . However, relying on examples in field theory, the authors of [9] suggest that the information that seem to get lost in the infinite momentum frame can actually be retrieved from (16). The Matrix Theory proposal is therefore that the action (16) fully represents M theory in the large  $N$  limit. The missing dimension in the previous counting is to be retrieved from such a limit. In fact a system of  $N$  D0-branes, as the one described by (16), has  $p_{11}^{t\sigma t} = \frac{N}{R}$  and the decompactification limit  $R \rightarrow \infty$  requires  $N \rightarrow \infty$  in some appropriate way.

That Matrix Theory represents M theory has been confirmed in a number of ways. Although there is still a long way to go to transform Matrix Theory into an efficient tool for computations, no contradiction has been found thus far with this hypothesis. MST can be thought of precisely as one of the most important confirmation of Matrix Theory. MST, first proposed in [13], is obtained in the following way. One first compactifies Matrix Theory along the 9th direction on a circle of radius  $R_9$  and ends up with a SYM theory in 1+1 dimensions. This operation is technically rather complicated, see [12], and will not be repeated here. After this operation one is left with two compactification radii,  $R = R_{11}$  along the 11th direction and  $R_9$ . One then exchanges the 9th and 11th direction. In the Matrix Theory limit  $R \rightarrow \infty$ , one recovers the appropriate setting for IIA theory: ten uncompactified dimensions plus one compactified along the circle of radius  $R_9$ , which has become the 11th direction of M theory, [13, 16]; it is this radius which is expected to correspond to  $g_s$ . *This is what is called MST. It is expected to represent a nonperturbative version of type IIA theory.*

## 4 Matrix String Theory

### 4.1 Minkowski version

The MST is defined by the following  $U(N)$  SYM model in a 1+1 Minkowski space, specified by the action

$$S = -\frac{1}{2\pi} \int d\sigma d\tau \text{Tr} \left( D_\mu X^i D^\mu X^i + \frac{1}{2g^2} F_{\mu\nu} F^{\mu\nu} - \frac{g^2}{2} [X^i, X^j]^2 - i\bar{\theta} \rho^\mu D_\mu \theta - g\theta^T \Gamma_i [X^i, \theta] \right), \quad (18)$$

where  $g$  is the gauge coupling,  $\sigma$  and  $\tau$  are the world-sheet coordinate on the cylinder.  $X^i$  with  $i = 1, \dots, 8$  are hermitean  $N \times N$  matrices and  $D_\mu X^i = \partial_\mu X^i + i[A_\mu, X^i]$ .  $F_{\mu\nu}$  is the curvature of  $A_\mu$ .  $\theta$  represents  $2N \times N$  matrices whose entries are simultaneously 2D and  $SO(8)$  spinors. It can be written as  $\theta^T = (\theta_s, \theta_c)$ , where  $\pm$  denotes the 2D chirality and  $\theta_s, \theta_c$  are spinors in the  $8_s$  and  $8_c$  representations of  $SO(8)$ , while  $^T$  represents the 2D transposition.  $\rho_\mu$  are again the 2D gamma matrices:  $\{\rho_\mu, \rho_\nu\} = -2\eta_{\mu\nu}$ , and  $\bar{\theta} = \theta^T \rho^0$ . The matrices  $\Gamma_i$  are the  $16 \times 16$   $SO(8)$  gamma matrices. The remaining conventions are as in (15).

For definiteness we will write the matrices  $\rho_\mu$  and  $\Gamma_i$  in the form

$$\rho^0 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \rho^1 = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}, \quad \Gamma_i = \begin{pmatrix} 0 & \gamma_i \\ \tilde{\gamma}_i & 0 \end{pmatrix}, \quad (19)$$

and  $\gamma_i, \tilde{\gamma}_i$  are the same as in Appendix 5B of [7].

The action (18) is invariant under the supersymmetric transformations

$$\begin{aligned} \delta X^i &= \frac{i}{g} \epsilon \Gamma^i \theta \\ \delta \theta &= \frac{1}{2g^2} \rho^{\mu\nu} F_{\mu\nu} \epsilon - \frac{i}{2} [X^i, X^j] \Gamma_{ij} \epsilon - \frac{1}{g} \rho^\mu D_\mu X^i \rho^0 \Gamma_i \epsilon \\ \delta A_\mu &= -i\bar{\epsilon} \rho_\mu \theta, \end{aligned} \quad (20)$$

where  $\epsilon^T = (\epsilon_s, \epsilon_c)$  are 8+8 transformation parameters.

## 4.2 Euclidean version

We make a Wick rotation and introduce the complex coordinates

$$w = \frac{1}{2}(\tau + i\sigma), \quad \bar{w} = \frac{1}{2}(\tau - i\sigma), \quad A_w = A_0 - iA_1, \quad A_{\bar{w}} = A_0 + iA_1.$$

The action becomes

$$S = \frac{1}{\pi} \int_C d^2w \operatorname{Tr} \left( D_w X^i D_{\bar{w}} X^i - \frac{1}{4g^2} F_{w\bar{w}}^2 - \frac{g^2}{2} [X^i, X^j]^2 \right. \\ \left. + i(\theta_s D_{\bar{w}} \theta_s + \theta_c D_w \theta_c) + ig\theta^T \Gamma_i [X^i, \theta] \right), \quad (21)$$

where  $C$  is the infinite Euclidean cylinder spanned by  $w$ . The supersymmetry transformations take the form

$$\begin{aligned} \delta X^i &= \frac{i}{g} (\epsilon_s^- \gamma^i \theta_c + \epsilon_c^+ \tilde{\gamma}^i \theta_s) \\ \delta \theta_s &= \left( -\frac{i}{2g^2} F_{w\bar{w}} + \frac{1}{2} [X^i, X^j] \gamma_{ij} \right) \epsilon_s^- - \frac{1}{g} D_w X^i \gamma_i \epsilon_c^+ \\ \delta \theta_c &= \left( \frac{i}{2g^2} F_{w\bar{w}} + \frac{1}{2} [X^i, X^j] \tilde{\gamma}_{ij} \right) \epsilon_c^+ - \frac{1}{g} D_{\bar{w}} X^i \tilde{\gamma}_i \epsilon_s^- \\ \delta A_w &= -2\epsilon_s^- \theta_s, \quad \delta A_{\bar{w}} = -2\epsilon_c^+ \theta_c, \end{aligned} \quad (22)$$

where

$$\gamma_{ij} = \frac{1}{2} (\gamma_i \tilde{\gamma}_j - \gamma_j \tilde{\gamma}_i), \quad \tilde{\gamma}_{ij} = \frac{1}{2} (\tilde{\gamma}_i \gamma_j - \tilde{\gamma}_j \gamma_i).$$

A string interpretation, which is what we want to arrive at, is more natural after the coordinate transformation, already considered above,  $w \rightarrow z = e^w$ , i.e. after passing from the cylinder to the complex plane with the origin deleted, i.e.  $\mathbb{C}^*$ , or the Riemann sphere  $\mathbb{CP}^1$  with two punctures.

## 4.3 The string interpretation

Let us see how a simple string interpretation arises if we take a naive strong coupling limit in (21). We rescale  $A \rightarrow \frac{1}{g}A$  and imagine the fields represent small oscillations about a trivial background. The naive strong coupling limit ( $g \rightarrow \infty$ ) in the action tells us that all fields commute, therefore they can be simultaneously diagonalized. The formal aspect of the resulting theory is that of the Green-Schwarz superstring theory in the light-cone approach, (15); however there is a significant variant. The fields are the eigenvalues of  $X_i$  and  $\theta$ . Therefore for each field of the Green-Schwarz theory we have here  $N$  fields. We will see later on a completely satisfactory interpretation of this fact. For the time being let us notice that the theory becomes a free theory of the diagonal degrees of freedom. The gauge freedom can be completely fixed up to a residual gauge invariance which takes the form of the Weyl group, i.e. the permutation group of  $N$  elements. We are therefore allowed to fix the boundary conditions for the diagonal degrees of freedom up to this residual gauge transformation. On this basis these degrees of freedom lend themselves naturally to an interpretation as free strings of various lengths. For example, let the eigenvalues of  $X^i$  be  $\hat{X}^i = \operatorname{Diag}(x_1^i, \dots, x_N^i)$  and let us consider the effect on such configuration of the element

$$\mathcal{P} = \begin{pmatrix} 0 & 0 & \dots & \dots & 1 \\ 1 & 0 & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix} \quad (23)$$

of the Weyl group. The boundary condition  $\hat{X}^i(2\pi) = \mathcal{P}\hat{X}^i(0)\mathcal{P}^{-1}$  implies that  $x_k^i(2\pi) = x_{k-1}^i(0)$ , and so the  $x_k^i$  form a unique long string of length  $2\pi N$ . Of course there are many other possible configurations, beside the one determined by (23), with strings of any length, each corresponding to a conjugacy class of the Weyl group, [16].

The length of a string, in the above sense, has a precise physical meaning. Recall that the correspondence with the Green-Schwarz theory is established in the light-cone gauge. In this framework the length of a string is identified with the  $p^+ = p^0 + p^9$  component of the string center of mass momentum, where 0, 9 are the time and longitudinal direction, which do not appear explicitly in (18).

The naive strong coupling limit of (18) is therefore a free theory of closed superstring of various lengths. This will be confirmed by a more careful analysis later on. For the time being we note that, if the limit we have described is the true strong coupling limit, we can interpret it as the weak coupling limit of (type II) string theory, i.e. if  $g_s$  is the string coupling,  $g_s \sim g^{-1}$ , [10, 13].

## 5 Riemannian Instantons and their construction

The string interpretation of the previous subsection will be confirmed by an analysis of the theory in the background of its instantons. Let us, preliminarily, describe such classical (supersymmetric) solutions. In fact we will look for classical Euclidean supersymmetric configurations that preserve half supersymmetry. To this end we set  $\theta = 0$  and look for solutions of the equations  $\delta\theta^\pm = 0$ , i.e. from (22),

$$\begin{aligned} \left(\frac{i}{2g^2}F_{w\bar{w}} + \frac{1}{2}[X^i, X^j]\tilde{\gamma}_{ij}\right)\epsilon_c^+ &= 0, \quad D_w X^i \gamma_i \epsilon_c^+ = 0 \\ \left(-\frac{i}{2g^2}F_{w\bar{w}} + \frac{1}{2}[X^i, X^j]\gamma_{ij}\right)\epsilon_s^- &= 0, \quad D_{\bar{w}} X^i \tilde{\gamma}_i \epsilon_s^- = 0 \dots \end{aligned} \quad (24)$$

Solutions of these equations that preserve half supersymmetry are the following ones. Set  $X^i = 0$  for all  $i$  except two, for definiteness  $X^i \neq 0$  for  $i = 1, 2$ ; remark that  $\gamma_{12}$  is an antisymmetric  $8 \times 8$  matrix, and  $\gamma_{12}^2 = -1$  and therefore its eigenvalues are  $\pm i$  (moreover  $\tilde{\gamma}_{12} = \gamma_{12}$ ). It is easy to show that there exists  $\epsilon^+$  and  $\epsilon^-$ , each with four independent components, such that

$$\gamma_{12}\epsilon^\pm = \pm i\epsilon^\pm, \quad \gamma_1\epsilon^+ = -i\gamma_2\epsilon^+, \quad \tilde{\gamma}_1\epsilon^- = i\tilde{\gamma}_2\epsilon^-.$$

Now it is convenient to introduce the complex notation  $X = X^1 - iX^2$ ,  $\bar{X} = X^1 + iX^2 = X^\dagger$ . Then the conditions to be satisfied in order to preserve half supersymmetry are

$$F_{w\bar{w}} - ig^2[X, \bar{X}] = 0 \quad (25)$$

$$D_{\bar{w}}X = 0, \quad D_w\bar{X} = 0. \quad (26)$$

These are the same equations as (2), but they have been obtained in the context of MST. The solutions of these equations preserve half supersymmetry. For this reason they are also referred to as *BPS instantons*.

It is easy to verify that, if non-trivial solutions to such equations exist, they satisfy the equations of motion of the action (21). The normalized<sup>1</sup> instanton action is the action with  $\theta = 0$ ,  $X^i = 0$  for  $i = 3, \dots, 8$

$$S_{inst} = \frac{1}{2\pi} \int d^2w \operatorname{Tr} \left( -X D_w D_{\bar{w}} \bar{X} - \bar{X} D_w D_{\bar{w}} X - \frac{1}{2g^2} F_{w\bar{w}}^2 + \frac{g^2}{2} [X, \bar{X}]^2 \right). \quad (27)$$

It is elementary to prove that  $S_{inst}$  vanishes in correspondence with smooth solutions of (25, 26).

From a mathematical viewpoint, (25, 26) are easily seen to identify a *Hitchin system* [28, 29] on a sphere with two punctures. In such systems,  $F$  is the gauge curvature in reference to a gauge

<sup>1</sup>The action (27) is normalized so as to avoid boundary terms due to partial integration.

vector bundle  $V$ , and  $X$  is the holomorphic section of the bundle  $\text{End}V \otimes K$ , where  $K$  is the canonical line bundle over the base (which is trivial in our case). Hitchin systems have appeared both in the mathematical and the physical literature, [30, 32, 31, 33].

One first remarkable property of the Hitchin systems is that they form an integrable system of equations. To see this let us show that they satisfy a zero curvature condition. Let us define the spectral connection

$$\mathcal{A}_w = A_w + \lambda g X, \quad \mathcal{A}_{\bar{w}} = A_{\bar{w}} - \frac{g}{\lambda} \bar{X}, \quad (28)$$

where  $\lambda$  is a spectral parameter. We can rewrite (2) as the zero curvature condition for the connection  $\mathcal{A}_w$

$$\begin{aligned} \mathcal{F}_{\bar{w}w} &= \partial_{\bar{w}} \mathcal{A}_w - \partial_w \mathcal{A}_{\bar{w}} + i[\mathcal{A}_{\bar{w}}, \mathcal{A}_w] \\ &= (F_{\bar{w}w} + ig^2[X, \bar{X}]) + \lambda g(D_{\bar{w}}X) - \frac{g}{\lambda}(D_w \bar{X}) = 0, \end{aligned}$$

for generic values of the spectral parameter.

### 5.1 General ansatz for Riemannian instantons

In order to find solutions  $(A, X)$  of (25,26), we follow and generalize the example  $N = 2$  presented in section 2. We start from the analogous ansatz

$$A_w = i\partial_w Y^\dagger (Y^{-1})^\dagger, \quad X = Y^{-1}MY \quad (29)$$

where  $Y$  is a generic element in the complex group  $SL(N, \mathbb{C})$  and  $M$  specifies a branched covering of the cylinder. As a consequence of (29) the equation  $D_{\bar{w}}X = 0$  is equivalent to  $\partial_{\bar{w}}M = 0$  or

$$\partial_{\bar{z}}M = 0. \quad (30)$$

With this parametrization, the spectral connection becomes

$$A_w = i\partial_w Y^\dagger (Y^{-1})^\dagger + \lambda g Y^{-1}MY, \quad A_{\bar{w}} = -iY^{-1}\partial_{\bar{w}}Y - \frac{g}{\lambda}Y^\dagger M^\dagger (Y^\dagger)^{-1}. \quad (31)$$

from which it is easy to extract the zero curvature equation and see that it is written in terms of  $YY^\dagger$  only.

This parametrization is defined therefore in terms of two factors  $Y$  and  $M$ . As above,  $Y$  will be referred to as the *group theoretical factor*, while  $M$  defines a general branched covering of the cylinder. The factor  $Y$  will be discussed below, while branched coverings will be discussed at length in section 7. For the time being let us give some essential information. Let us consider the polynomial

$$P_X(y) = \text{Det}(y - X) = y^N + \sum_{i=0}^{N-1} y^i a_i,$$

where  $y$  is a complex indeterminate. The equation

$$P_X(y) = 0 \quad (32)$$

can also be written as the matrix equation

$$X^N + a_{N-1}X^{N-1} + \dots + a_0 = 0. \quad (33)$$

A diagonalizable matrix, which is solution of eq. (33), can always be cast in the canonical form

$$M = \begin{pmatrix} -a_{N-1} & -a_{N-2} & \dots & \dots & -a_0 \\ 1 & 0 & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix}. \quad (34)$$

The case considered in section 2 is evidently a particular case of the above formulas.

Due to (30), we have  $\partial_z a_i = 0$ , which means that the set of functions  $\{a_i\}$  are analytic in the complex plane, although they are allowed to have poles at  $z = 0$  and  $z = \infty$ . The point is that, as we shall see in many examples, Eq.(32) identifies in the  $(z, y)$  space a Riemann surface  $\Sigma$ , which is an  $N$ -sheeted branched covering of the cylinder. Generalizing what has been said in section 2, the explicit form of the covering is given by the set  $\{x^{(1)}(z), \dots, x^{(N)}(z)\}$  of eigenvalues of  $X$ . Each eigenvalue spans a sheet. The projection map to the base cylinder  $\mathcal{C}$  will be denoted  $\pi : \Sigma \rightarrow \mathcal{C}$ . The points where two or more eigenvalues coincide are called branch points. The identification cuts in the sheets start or end at these points (which may include 0 and  $\infty$ ). We stress that the covering is independent of the coupling  $g$ .

The Riemann surface  $\Sigma$  will be characterized by the genus  $h$  and by a certain number  $n$  of punctures. Each puncture will come associated with an integer, the length of the string entering or exiting at that puncture. Moreover each surface is characterized by certain numbers, the moduli. All this information is stored in the  $a_i$  analytic functions and will be worked out in sections 7 and 8.

## 5.2 Construction of instanton solutions

Each solution of (25), (26) consists of two parts: a branched covering of the cylinder via the relative  $X$  characteristic polynomial and a group theoretical factor. The aim of the present subsection is to generalize the example of section 2, by constructing the instanton solutions corresponding to the most general covering.

Let us recall our ansatz (29). The group theoretical factor  $Y$  takes values in the complex group  $SL(N, \mathbb{C})$ , while the matrix  $M$  determines the branched covering. The dependence on the Yang-Mills coupling constant  $g$  is contained in the  $Y$  factor, while  $M$  does not depend on  $g$ . In section 2 we have shown an example in which  $Y = KL$  where  $L$ , the *dressing factor*, tends to 1 in the strong coupling limit outside the string interaction points, while  $K$  is a special matrix, independent of  $g$ , endowed with the property that  $K^{-1}MK$  and  $K^\dagger M^\dagger (K^\dagger)^{-1}$  are simultaneously diagonalizable.

We will proceed in a parallel way also in the general case. It is well-known, [17], that the matrix  $M$  can be diagonalized

$$M = S \hat{M} S^{-1}, \quad \hat{M} = \text{Diag}(\lambda_1, \dots, \lambda_N) \quad (35)$$

by means of the following matrix  $S \in SL(N, \mathbb{C})$ :

$$S = \Delta^{-\frac{1}{N}} \begin{pmatrix} \lambda_1^{N-1} & \lambda_2^{N-1} & \dots & \dots & \lambda_N^{N-1} \\ \lambda_1^{N-2} & \lambda_2^{N-2} & \dots & \dots & \lambda_N^{N-2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 1 & \dots & \dots & 1 \end{pmatrix}, \quad (36)$$

where

$$\Delta = \prod_{1 \leq i < j \leq N} (\lambda_i - \lambda_j). \quad (37)$$

We notice first that the role of  $a$  in section 2 is played now by  $\Delta$ , which vanishes whenever two eigenvalues coincide. Two coinciding eigenvalues define a branch point of the covering. We will return later on to this issue. For the time being let us notice that, like in section 2, going around a branch point in the complex  $z$ -plane, produces a reshuffling of the eigenvalues that can be represented via a monodromy matrix  $\Lambda: \hat{M} \rightarrow \Lambda \hat{M} \Lambda^{-1}$ . Correspondingly we have  $S \rightarrow S \Lambda^{-1}$ , so that the single-valuedness of  $M$  is preserved.

The construction of  $K$  and  $L$  in the general case is subtle and technically rather complicated. The explicit construction of  $K$  and  $L$  are given in [21] and will not be reported here. However, on the basis of the example of section 2, it is not difficult to understand the general features of these matrices.



First one introduces a monodromy-invariant  $K$  such that  $K^{-1}S = U$  be unitary. As it turns out,  $K$  may have singularities at the points of  $\mathcal{C}$  where any two eigenvalues of  $M$  coincide, i.e. at the branch points of the spectral covering (the elements of  $K$  contains as factors fractional powers of  $|\Delta|$ , much in the same way as  $K$  in (5) contains fractional powers of  $|a|$ ). Therefore  $K^{-1}MK$  is in general singular at these points. That is why we must introduce into the game a new monodromy invariant matrix  $L$ , with the purpose of canceling the singularities of  $K^{-1}MK$  in such a way that  $L^{-1}K^{-1}MKL$  be smooth and satisfy (25), (26). The entries of  $L$  can be taken to be generalizations of the  $u$  field of section 2. Let us denote again by  $u$  any such field. For (25), (26) to be satisfied these fields  $u$  must satisfy, [21], an equation of the WZNW type with the following general structure

$$\partial_w \partial_{\bar{w}} u + \dots \sim \partial_w \partial_{\bar{w}} \ln |\Delta| = \pi \frac{\partial \Delta}{\partial w} \frac{\partial \bar{\Delta}}{\partial \bar{w}} \delta(\Delta), \quad (38)$$

where dots represent all the other terms, which are irrelevant in the cancellation of singularities. Let us refer to these equations as the ‘dressing equations’. On the right-hand side we see the typical delta-function-type source which characterizes them. The sources are point-like and located at the zeroes of  $\Delta$ , that is at the branch points of the covering.

By construction  $K$  is independent of  $g$  while  $L$  does depend on  $g$ . One can show that in fact  $L \rightarrow 1$  as  $g \rightarrow \infty$ . Let us present a simple argument in this sense.

The solution  $X$  exists with the required properties only if the ‘dressing equations’ admit solutions that vanish at  $w = \pm\infty$ . To our best knowledge, not much is known in the literature concerning the existence of such solutions. Based on the analysis of [20], we assume that the ‘dressing equations’ do admit solutions that vanish at  $w = \pm\infty$ . Once one assumes this, it is rather easy to argue, on a completely general ground, that in the strong coupling limit,  $g \rightarrow \infty$ , such solutions vanish outside the zeroes of the discriminant. The argument goes as follows. Consider a candidate solution of (25) in which  $u = 0$  outside the zeroes of the discriminant, for all the  $u$ ’s. Then, there,  $L = 1$ , and  $X = K^{-1}MK$ . As noted previously, in such a situation  $[X, \bar{X}] = 0$ , since both  $X$  and  $\bar{X}$  are simultaneously diagonalized by the matrix  $U = K^{-1}S$ . Now we have to show that also  $F_{w\bar{w}}$  vanishes outside the zeroes of the discriminant if  $L = 1$ . In fact when  $L = 1$ ,

$$A_{\bar{w}} = -iK^{-1}\partial_{\bar{w}}K = -i(K^{-1}SS^{-1})\partial_{\bar{w}}(SS^{-1}K) = -iU(\partial_{\bar{w}} + \bar{A}_{\bar{w}})U^{-1},$$

where  $\bar{A}_{\bar{w}} = S^{-1}\partial_{\bar{w}}S$ . But  $\partial_{\bar{w}}S \equiv 0$  due to holomorphicity of the eigenvalues of  $M$ . In conclusion (25) is identically satisfied by the ansatz  $L = 1$  outside the zeroes of the discriminant. Since the solutions are uniquely determined by their boundary conditions, we can conclude that, as  $g \rightarrow \infty$ , the only solution of the dressing equations outside the zeroes of the discriminant, is the identically vanishing solution. We infer from this argument that the solutions of the dressing equations for large  $g$  are concentrated around the branch points and become more and more spiky as  $g$  grows larger and larger. Therefore the matrix  $L$  has the properties we expect.

The previous argument hinges on the occurrence that as  $g = \infty$  we have both  $[X, \bar{X}] = 0$  and  $F_{w\bar{w}} = 0$ . Other types of solutions can be envisaged, see [21]. However here we limit ourselves precisely to those solutions of the Hitchin equations (25), (26) for which at  $g = \infty$  we have both  $[X, \bar{X}] = 0$  and  $F_{w\bar{w}} = 0$  and call them *stringy instantons*, since they represent interpolating solutions between genuine initial and final string configurations.

To end this section let us emphasize the double ‘miracle’ of the above construction: we construct an everywhere smooth solution by means of two non smooth matrices  $K$  and  $L$ , which are such that on the one hand  $L \rightarrow 1$  as  $g \rightarrow \infty$  and on the other hand  $K$  form together with  $S$  a unitary matrix  $U = K^{-1}S$ , so that as  $g \rightarrow \infty$ ,  $X \rightarrow U\hat{M}U^{-1}$ . This will be adequately appreciated in the next section.

### 5.3 Summary

In this section we have seen that *utilizing a very general recipe, an instanton solution can be constructed in correspondence with the most general branched covering of the base cylinder, i.e. to*

any Riemann surface with punctures that can be represented as a branched covering of the cylinder we can associate an instanton. We have of course still to clarify how general these Riemann surfaces are. This requires a more explicit analysis of branched coverings. At this point, however, before we do that, we believe it more useful to first see how stringy instantons can be used to compute physical processes. This will clarify the physical setup of our discussion and what branched coverings we really have to describe and moreover it will introduce a new input in the discussion: the moduli space of stringy instantons.

## 6 MST on a Riemannian instanton background

The aim of this section is to show how the above construction of stringy instantons can be transformed into a calculational tool. In particular we would like to interpret an amplitude on a given instanton background as an amplitude for a string process characterized by the Riemann surface which underlies the instanton.

Below we show that MST in the strong coupling limit in the background of a given stringy instanton solution reduces to the Green-Schwarz superstring theory plus a decoupled Maxwell theory, and compute string interaction amplitudes in such background. Since the latter interpolates between an initial and a final string configuration via a punctured Riemann surface  $\Sigma$  (which represents a branched covering of the base cylinder), the amplitudes can be interpreted in a string theoretic way as the transition amplitudes between two such configurations. The purpose of this section is to show that their leading term is proportional to  $g_s^{-\chi}$ , where  $\chi = 2 - 2h - n$  is the Euler characteristic of the Riemann surface of genus  $h$  with  $n$  punctures, which characterizes the given classical solution. This is the result one expects from perturbative string interaction theory and proves beyond any doubt the relation between MST at strong coupling and string interaction theory.

### 6.1 Expansion about a Riemannian instanton

What we want to do is to split every field  $\Phi$  in the action (21) into a classical instanton background part and a quantum fluctuating part

$$\Phi = \Phi^{(b)} + \phi \quad (39)$$

and expand the action. The instanton action vanishes, as we have seen above, and the piece of action linear in  $\phi$  vanishes due to the equations of motion. We want to evaluate what remains in the strong coupling limit, by considering an expansion of the action in  $1/g$ .

As a first step let us analyze the background part. To this end let us recall a few facts from the previous section. The dependence on the coupling is entirely contained in the factor  $L$ . We have seen that in the strong coupling limit  $L \rightarrow 1$  outside the branch points of the covering. Since here we are interested in expanding the action (21) in inverse powers of  $1/g$ , and actually in singling out the dominant term in this expansion (see below), we will consider the action (21) around a given classical solution stripped of the above dressing factor, and exclude from the integration region the branch points on the cylinder. In other words we will consider from now on the action (21) in which the relevant  $Y$  is replaced by  $K$  and the integral extends over  $C_0$  which is the initial cylinder  $\mathcal{C}$  from which small disks have been cut out around the branch points. Said otherwise, we introduce in our integrated action a regulator (which will eventually be removed).

After getting rid of the dressing factor, the classical background configuration is specified by  $X = K^{-1}MK$  and  $A_{\bar{w}} = -iK^{-1}\partial_{\bar{w}}K$ . As expected, this configuration is singular exactly at the branch points. We have seen that  $M = \hat{M}S^{-1}$ .  $\hat{M}$  is the matrix of eigenvalues of  $M$  and of  $X$ , so we denote it equivalently by  $\hat{X}$ . In the strong coupling limit  $X \rightarrow U\hat{X}U^{-1}$ , where  $U = K^{-1}S$  is a unitary matrix and therefore simultaneously diagonalizes  $X$  and  $\hat{X}$ .  $U$  is singular at the branch points both in the sense that it may diverge there and that it undergoes a monodromy transformation upon going around a branch point. Corresponding to  $\hat{X}$  we have  $\hat{A}_w$  which, as

was shown above, vanishes everywhere, even at the branch points. What has happened is that the unitary transformation has swallowed entirely the connection, including the singularities.

$U$  is singular at the branch points, but finite (and multi-valued) in  $\mathcal{C}_0$ . Therefore, with a gauge transformation, we can remove the unitary factor  $U$  from the action defined in  $\mathcal{C}_0$ . This leads us to the

- diagonal representation:  $\hat{X}$  diagonal and  $\hat{A} = 0$ .

for the classical background in the strong coupling limit.

A comment about our use of singular gauge transformations is in order. As we have stressed several times, the classical configuration specified by a given couple  $(X, A)$  is smooth on the initial cylinder  $\mathcal{C}$ , but if we strip the solution of the dressing factor we get a configuration which is singular exactly at the branch points. The dressing factor is there exactly to compensate for these singularities.  $L - 1$  has support only at the branch points in the strong coupling limit. Therefore we replace  $\mathcal{C}$  with  $\mathcal{C}_0$  by excluding the branch points from the integration region to preserve smoothness. This is justified by the following consideration. Beside the initial smooth configuration  $(X, A)$  on  $\mathcal{C}$ , we will meet another smooth situation when we lift our theory to the branched covering  $\Sigma$  of  $\mathcal{C}$  (see below). At that stage the regulator can be removed and the branch points can be restored in the integration region. What has happened is that, in order to pass to the covering, we need a singular gauge transformation, the  $U$  used above, which exactly kills the singularity exposed by the strong coupling limit. The transformation  $U$  is not single-valued, since it picks up a phase when crossing a cut. In other words, it is natural to perform this singular gauge transformation if we want to reach a smooth situation which is fit for field theory.

Let us return now to the action (21) with the above understanding of the background part. To extract the strong coupling effective theory, we first rewrite the action in the following useful form

$$S = \frac{1}{\pi} \int d^2w \operatorname{Tr} \left( D_w X^I D_{\bar{w}} X^I - \frac{g^2}{2} [X^I, X^J]^2 - g^2 [X^I, X] [X^I, \bar{X}] - \frac{1}{2} \bar{X} D_w D_{\bar{w}} X \right. \\ \left. - \frac{1}{2} X D_w D_{\bar{w}} \bar{X} - \frac{1}{4g^2} (F_{w\bar{w}} - ig^2 [X, \bar{X}])^2 + i(\theta_s D_{\bar{w}} \theta_s + \theta_c D_w \theta_c) + ig \theta^T \Gamma_i [X^i, \theta] \right),$$

where  $I = 3, 4, \dots, 8$ . We now expand the action around a generic instanton configuration as in (39), but we further split the quantum field  $\phi$

$$\Phi = \Phi^{(b)} + \phi^t + \phi^n \equiv \Phi^{(b)} + \phi \equiv \Phi^o + \phi^n, \quad (40)$$

where  $\Phi^{(b)}$  is the background value of the field at infinite coupling,  $\phi^t$  are the fluctuations along the Cartan directions and  $\phi^n$  are the fluctuations along the complementary directions in Lie algebra  $\mathfrak{u}(N)$ . Of course only the upper case fields  $X, \bar{X}$  and  $A$  will have in general non zero background value. In the following we suppose we have carried out the operation described above and by background value we refer to the diagonal representation.

As we have already noticed, the expansion of the action starts with quadratic terms in the fluctuations.

## 6.2 Fixing the gauge and integrating along the non-Cartan directions

To proceed further let us fix the gauge. We use, in the strong coupling limit, the following gauge, inspired by the one used in [35]

$$\mathcal{G}_{w\bar{w}} = D_w^o a_{\bar{w}} + D_{\bar{w}}^o a_w + ig^2 ([X^o, \bar{x}] + [\bar{X}^o, x]) + 2ig^2 [X^{oI}, x^I] = 0, \quad (41)$$

where  $D^o$  is the covariant derivative with respect to  $A^o$ . Next we introduce the Faddeev-Popov ghost and antighost fields  $c$  and  $\bar{c}$  and expand them like all the other fields. Then we add to the action the gauge fixing term

$$S_{gf} = \frac{1}{4\pi g^2} \int d^2w \mathcal{G}_{w\bar{w}}^2 \quad (42)$$

and the corresponding Faddeev-Popov ghost term

$$S_{ghost} = -\frac{1}{2\pi g^2} \int d^2w \bar{c} \frac{\delta \mathcal{G}_{w\bar{w}}}{\delta c} c, \quad (43)$$

where  $\delta$  represents the gauge transformation with parameter  $c$ .

At this point, to single out the strong coupling limit of the action, we rescale the fields in appropriate manner. Precisely, we redefine our fields as follows

$$A_w = g a_w^t + a_w^n, \quad X = \hat{X} + x^t + \frac{1}{g} x^n, \quad X^I = x^{I^t} + \frac{1}{g} x^{I^n}, \quad \theta^n = \theta^t + \frac{1}{\sqrt{g}} \theta^n$$

and likewise for the conjugate variables. For the ghosts we set

$$c = g c^t + \sqrt{g} c^n, \quad \bar{c} = g \bar{c}^t + \frac{1}{\sqrt{g}} \bar{c}^n.$$

One may wonder why we choose such rescalings and not others. A partial answer is in the remark that these rescalings introduce a unit Jacobian in the path integral measure of the non-zero modes, although they may produce a non-trivial factor due to the presence of zero modes (see below). However this is not a sufficient criterion to completely fix the rescalings. Let us say that this operation leads us to a sensible strong coupling limit and we can take it as our definition of the strong coupling theory.

After these rescalings the action becomes

$$S = S_{sc} + S_n + o\left(\frac{1}{\sqrt{g}}\right),$$

where

$$S_{sc} = \frac{1}{\pi} \int_{C_0} d^2w \text{Tr} \left[ d_w x^{I^t} d_{\bar{w}} x^{I^t} + d_w x^t d_{\bar{w}} \bar{x}^t + i(\theta_s^t d_{\bar{w}} \theta_s^t + \theta_c^t d_w \theta_c^t) \right. \\ \left. + d_w a_{\bar{w}}^t d_{\bar{w}} a_w^t + d_w \bar{c}^t d_{\bar{w}} c^t \right] \quad (44)$$

$S_n$  is the purely quadratic term in the  $\phi^n$  fluctuations. The latter can be easily integrated over and, since they do not involve zero modes, give exactly 1. Let us see this in detail.

$S_n$  has the form

$$S_n = \frac{1}{\pi} \int d^2w \text{Tr} \left[ \bar{x}^n Q x^n + x^{I^n} Q x^{I^n} + a_{\bar{w}}^n Q a_w^n + \bar{c}^n Q c^n + i(\theta_s^n, \theta_c^n) \mathcal{A} \begin{pmatrix} \theta_s^n \\ \theta_c^n \end{pmatrix} \right], \quad (45)$$

where

$$Q = \text{ad}_{\bar{X}^0} \cdot \text{ad}_{X^0} + \text{ad}_{a_{\bar{w}}^t} \cdot \text{ad}_{a_w^t} + \text{ad}_{x^{I^t}} \cdot \text{ad}_{x^{I^t}}$$

and

$$\mathcal{A} = \begin{pmatrix} i \text{ad}_{a_{\bar{w}}^t} & \gamma_i \text{ad}_{X^{0i}} \\ \tilde{\gamma}_i \text{ad}_{\bar{X}^{0i}} & i \text{ad}_{a_w^t} \end{pmatrix}.$$

In the path integral we can now integrate over the non-Cartan modes and obtain a ratio of determinants of  $\mathcal{A}$  and  $Q$ . Since these operators do not have zero modes the calculation is elementary. The integration over  $a^n$  and the conjugates exactly cancels the integration over  $c^n$  and the conjugates. What remains is a ratio  $((\text{Det} \mathcal{A})^{16} / (\text{Det} Q)^8)^{N^2 - N}$ . The expression of the numerator is formal: one should understand  $\text{Det} \mathcal{A}$  as  $\sqrt{\text{Det}(-\mathcal{A} \mathcal{A}^\dagger)}$ . But  $\mathcal{A} \mathcal{A}^\dagger = \mathcal{A}^\dagger \mathcal{A} = -Q$ . Therefore the net result of integrating over the non-Cartan modes is 1. This is the result expected from supersymmetry in the absence of zero modes.

In conclusion, in the strong coupling limit we are left with the quadratic action (44) over the Cartan modes.

### 6.3 Lifting to the branched covering

Let us now show that the effective theory we obtained in the previous subsection corresponds to the Green–Schwarz superstring plus a free Maxwell action on the worldsheet identified by the branched covering of the relevant background. To this end we recall the quadratic action (44). Since all the matrices involved are diagonal we can rewrite this action in terms of the diagonal modes  $\phi^i = \phi_{(1)}, \dots, \phi_{(N)}$ :

$$S_{sc} = \frac{1}{\pi} \int_{\mathcal{C}_0} d^2w \sum_{n=1}^N \left[ \partial_w x_{(n)}^i \partial_{\bar{w}} x_{(n)}^i + i(\theta_{s(n)} \partial_{\bar{w}} \theta_{s(n)} + \theta_{c(n)} \partial_w \theta_{c(n)}) \right. \\ \left. + \partial_w a_{\bar{w}(n)} \partial_{\bar{w}} a_{w(n)} + \partial_w \bar{c}_{(n)} \partial_{\bar{w}} c_{(n)} \right]. \quad (46)$$

This is a theory of free fields on  $\mathcal{C}_0$  and it is tempting to extend the action to  $\mathcal{C}$  by just forgetting the punctures on the cylinder corresponding to the branch points. However this is not correct. The fields  $x^i$  are not single-valued on the cylinder. For example, upon going around a simple branch point, at least one  $x^i$  is mapped to another one, and this is precisely the way a joining or a splitting of two strings is represented in this formalism, [18][19]. There are possibly many joining and splitting of strings in the process spanned by the instantons under consideration, and we can repeat the above word by word for any string interaction point. What is suggested here is that the fields  $x^i$  (as well as the others) are not well defined on each sheet, but all together they form a well defined field on the covering surface. Mathematically, this problem can be rephrased as follows: the fields in (46) are not section of bundles over  $\mathcal{C}$ ; however they can be combined to form sections of line bundles on the covering  $\Sigma$  of  $\mathcal{C}$ .

At this point it is worth spending a few words about *Hitchin systems*. The Hitchin systems we are interested in are defined starting from a  $U(N)$  vector bundle  $V$  over  $\mathcal{C}$ , associated with the fundamental representation of  $U(N)$ . They consist of couples  $(A, X)$  where  $A$  is a gauge connection and  $X$  a section of  $End V \otimes K$ , where  $K$  is the canonical bundle of  $\mathcal{C}$ , which satisfy (25) and (26), [28]. Such systems can be lifted to an  $N$ -branched covering of  $\mathcal{C}$ , [29], [30], [32]. A remarkable feature of the lifting is the appearance on the branched covering of a line bundle  $L$  constructed out of  $V$  and from which in turn  $V$  can be reconstructed. In simple words the initial non-Abelian system can be described by an equivalent Abelian system on the branched covering.

This is the same situation we are faced here. In fact let us look at the realization of local fields on a Riemann surface represented as a branched covering. If  $\Sigma$  is a branched covering of the cylinder  $\mathcal{C}$ , then, as we have seen, there is a projection map  $\pi : \Sigma \rightarrow \mathcal{C}$  whose inverse image is  $N$ -valued. In our language this is simply

$$\pi^{-1} : w \rightarrow (x_{(1)}(w), \dots, x_{(N)}(w)). \quad (47)$$

Suppose a local complex field  $\tilde{\psi}$  is given on  $\Sigma$ ; applying the above construction  $\tilde{\psi}$  can be represented as an  $N$ -tuple  $(\psi_{(1)}(w), \dots, \psi_{(N)}(w))$  representing the field on each copy of the cylinder  $\mathcal{C}$  that composes the covering  $\Sigma$ ; the  $\psi_{(i)}(w)$ 's are related by the appropriate monodromy properties along the cuts.

Going now back to the action (46), we see that we have to interpret any set of  $N$  fields  $(\phi_{(1)}, \dots, \phi_{(N)})$  in it as a unique field  $\tilde{\phi}$  on the covering  $\Sigma$ .  $\tilde{\phi}$  is locally a function of a coordinate  $z$  in  $\Sigma$ . From the point of view of  $\Sigma$ , the  $w$  coordinate is locally defined via an abelian differential  $\omega = dw$  with imaginary periods, which is canonical, i.e. is fixed only by the complex structure of the surface (see [39], [40] and below).

Finally we can write the strong coupling action (46) as follows

$$S_{sc}^{\Sigma} = S_{GS}^{\Sigma} + S_{Maxwell}^{\Sigma}, \quad (48)$$

$$S_{GS}^{\Sigma} = \frac{1}{\pi} \int_{\Sigma} d^2z \left( \partial_z \tilde{x}^i \partial_{\bar{z}} \tilde{x}^i + i(\tilde{\theta}_s \partial_{\bar{z}} \tilde{\theta}_s + \tilde{\theta}_c \partial_z \tilde{\theta}_c) \right) \quad (49)$$

$$S_{Maxwell}^{\Sigma} = \frac{1}{\pi} \int_{\Sigma} d^2z \left( g^{z\bar{z}} \partial_z \tilde{a}_{\bar{z}} \partial_{\bar{z}} \tilde{a}_z + \partial_z \tilde{c} \partial_{\bar{z}} \tilde{c} \right). \quad (50)$$

In (49) a  $\sqrt{\omega_z}$  (resp.  $\sqrt{\omega_{\bar{z}}}$ ) factor has been absorbed in  $\tilde{\theta}_s$  (resp.  $\tilde{\theta}_c$ ) which is a  $(\frac{1}{2}, 0)$  (resp.  $(0, \frac{1}{2})$ ) differential on  $\Sigma$  and the metric in the Maxwell term is  $g_{z\bar{z}} = \omega_z \omega_{\bar{z}}$ . In the integrals in (49) we have ignored the existence of small discs cut out around the branch points. This is allowed since everything now is smooth in  $\Sigma$  (we can remove the regulator introduced above).

Summarizing, what we obtained in this subsection is that the strong coupling effective theory is given by the Green-Schwarz superstring action on the branched covering worldsheet plus a decoupled Maxwell theory on the same surface. The fields in (49) are now sections of line bundles over  $\Sigma$ , i.e. they are well defined fields on the Riemann surface: for example  $\tilde{a}$  is a section of the canonical bundle of  $\Sigma$ , and so on.

## 6.4 String amplitudes

Let us compute first, for simplicity, the vacuum to vacuum amplitude of the SYM theory in the strong coupling limit in the background of a given instanton. As we have already pointed out several times, this amplitude (up to the vertex string insertions, see below) has a string interpretation as the amplitude for the transition from the initial to the final string configuration described by the instanton. If this interpretation is correct this amplitude, to the leading order, should be proportional to  $g_s^{-\chi}$  where  $\chi$  is the Euler characteristic of the Riemann surface  $\Sigma$ , i.e. the covering surface introduced above.

What remains for us to do in order to evaluate this amplitude is to integrate over the Cartan modes in the functional integral with action (49) (the non-Cartan modes have been integrated out above). Since the action is free, the integration produces a ratio of determinants, which turns out to be a constant. However we have to take account of the zero modes for the fields that have been rescaled (the unrescaled zero modes are irrelevant in this argument). The rescaled fields in  $\mathcal{C}$  are the Maxwell and the ghost fields. The corresponding fields in  $\Sigma$  will be rescaled too

$$\tilde{a}_z \rightarrow g \tilde{a}_z, \quad \tilde{a}_{\bar{z}} \rightarrow g \tilde{a}_{\bar{z}}, \quad \tilde{c} \rightarrow g \tilde{c}, \quad \tilde{\bar{c}} \rightarrow g \tilde{\bar{c}}. \quad (51)$$

Therefore let us single out the Maxwell (plus ghost) partition function. We will show that the decoupled  $U(1)$  theory is there to generate the stringy factor  $g_s^{-\chi}$  as a consequence of the rescaling (51)

In fact under this rescaling the Maxwell partition function ( $a = a_z, \bar{a} = a_{\bar{z}}$ )

$$Z_{\text{Maxwell}}^{\Sigma} = \int \mathcal{D}[\tilde{a}, \tilde{\bar{a}}, \tilde{c}, \tilde{\bar{c}}] e^{-S_{\text{Maxwell}}^{\Sigma}(\tilde{a}, \tilde{\bar{a}}, \tilde{c}, \tilde{\bar{c}})}$$

rescales with a factor depending on the zero modes. Roughly speaking, what happens is that the above integral is interpreted as the ratio  $(\text{Det}' \square_c) / (\text{Det}' \square_a)$ , where  $\square = \partial \bar{\partial}$  denotes the quadratic operator in the action, and ' means that the zero modes have been excluded from the computation of the regularized determinants; since we have rescaled the measure there will arise a factor of  $g$  to a power equal to the unbalance of the zero modes (a more precise account of this point can be found in [36]). The problem is therefore to count the latter. As for the ghost fields which are scalars, the only zero modes of the  $\bar{\partial}$  operator on  $\Sigma$  is the constant. The zero modes of the Maxwell fields correspond to the holomorphic differential on  $\Sigma$ . If  $\Sigma$  were a closed Riemann surface of genus  $h$ , their number would be  $h$ . Their counting in the present case is not completely standard as  $\Sigma$  is actually a Riemann surface with punctures (representing the in- and out- strings). For the purpose of such counting we can replace punctures with boundaries, since the Euler characteristic does not change. A way to do the counting is to construct the double  $\hat{\Sigma}$  of  $\Sigma$ :  $\hat{\Sigma}$  has genus  $\hat{h} = 2h + b - 1$ , where  $b$  is the number of boundaries, and admits an anticonformal involution with the set of fixed points corresponding exactly to the boundary of  $\Sigma$ . We can count now the number of analytic differential on  $\Sigma$  that extend to  $\hat{\Sigma}$ , that is the so-called analytic Schottky differentials [37]: their number is  $\hat{h}$ . Therefore the overall unbalance of zero modes (including the ghosts) is  $\hat{h} - 1 = 2h + b - 2$ , (or, equivalently,  $2h + n - 2$ , if  $n$  is the number of punctures, which is more appropriate in our case). This is exactly the opposite of the Euler number of  $\hat{\Sigma}$ . An equivalent

way of deriving this result is to use the Gauss-Bonnet theorem on  $\hat{\Sigma}$  and noting that, due to the involution, the integral of the curvature over  $\Sigma$  is one half of the total contribution.

Finally the factor in front of the vacuum to vacuum amplitude will be  $g^{-2h-n+2} = g_s^{2h+n-2}$ . The exponent of  $g$  is precisely the Euler characteristic of  $\Sigma$ , as we wanted to prove.

In order to appreciate exactly what we have just computed we must now specify what it corresponds to in string interaction theory. In this sense the amplitude we have just computed in the strong coupling limit is a basic amplitude but, of course, an incomplete one.

First of all, real string amplitudes should contain vertex string insertions, i.e. should be correlators of the vertex operators corresponding to the various in- and out- (super)strings. In this regard we simply remark that such vertex operators are constructed in terms of the string fields  $\tilde{x}^i$  and  $\tilde{\theta}$ , therefore the treatment of the non-Cartan modes above is not affected and the discussion of the zero modes of the Maxwell sector is unchanged. Therefore the scaling factor  $g_s^{-\chi}$  is left unchanged too.

Moreover, in order to obtain complete amplitudes, we must still integrate over the moduli of the Hitchin systems, i.e. over the inequivalent Riemannian instantons that interpolate between a given initial and a given final state. If we want to implement such more advanced stage of calculation, we have to take into account in the measure some Jacobians that are produced by the various field splittings we have considered above. Following [38], the background/fluctuations splitting of the fields in the path integral generates a Jacobian  $J_{b/f}$ . In an analogous way also the Cartan/non-Cartan splitting gives rise to another Jacobian factor  $J_{C/nC}$ . These factors are easily seen to depend only on the Cartan modes of  $X$  and  $\theta$ . This, in particular, implies the validity of the procedure we used for the integration over the non-Cartan modes.

So let us introduce the vertex operators  $V_1, \dots, V_n$  corresponding to  $n$  incoming and outgoing strings, expressed in terms of  $\tilde{x}, \tilde{\theta}$ , and of the string transverse momenta, and insert them into the path integral. The genus  $h$  amplitude (in the strong coupling limit) will schematically be:

$$\langle V_1, \dots, V_n \rangle_h = g_s^{-\chi} \int_{\mathcal{M}_N^{(h,n)}} dm \int \mathcal{D}[\tilde{x}, \tilde{\theta}, \tilde{a}, \tilde{c}] J_{b/f} J_{C/nC} V_1 \dots V_n e^{-S_{GS} - S_{Maxwell}}, \quad (52)$$

We have singled out the integration over  $\mathcal{M}_N^{(h,n)}$ , namely over all distinct instantons which underlie the given string process for fixed  $N$ , that is to say with assigned incoming and outgoing strings and string interactions. In ordinary string interaction theory  $\mathcal{M}^{(h,n)}$  is nothing but the moduli space of Riemann surfaces of genus  $h$  with  $n$  punctures, a complex space of dimension  $3h - 3 + n$ .

What actually  $\mathcal{M}_N^{(h,n)}$  is in MST is the main subject of the next sections.

However, before we pass to this subject, we owe one more comment on the Maxwell action. As we have already said, the role of  $S_{Maxwell}$  in (52) is to ensure the correct factor  $g_s^{-\chi}$  in front of the amplitude. For the rest its integration simply gives a number in front of the amplitude, since the Maxwell and ghost modes do not interact with the other modes.

## 6.5 Summary

In this section we have seen that *by expanding the MST action about a Riemannian instanton one gets, in the strong coupling limit, the Green-Schwarz action plus the free Maxwell action over the Riemann surface supporting the instanton. If this Riemann surface has genus  $h$  and  $n$  punctures, the path integral is proportional to a factor  $g^{-2h-n+2} = g_s^{2h+n-2}$ . This is the correct factor one expects from string interaction theory for a string process mediated by such a Riemann surface.* What remains now to examine is how general this result is compared to what is required by string interaction theory. To this end a closer analysis of branched coverings cannot be further postponed.

## 7 Riemann Surfaces as Plane Curves

It is time to turn to a more careful description of the second ingredient of stringy instantons, i.e. to branched coverings of the cylinder (or the Riemann sphere with two punctures). This section

as well as the following one is rather technical. By this I mean that the relevant mathematical concepts and terminology are not frequently met, for the time being, in the theoretical physics literature. However, even though I will try to keep a low lexical profile, a complete understanding of the problems we intend to cope with here requires the use of such concepts and terminology.

The purpose of this section is to explicitly show how the Riemann surfaces that are necessary in string interaction theory arise as branched coverings of the cylinder. To this end let us return to the definition in section 3.5, in particular eq.(32) or (33). Branched coverings make their appearance in MST as solutions of affine equations

$$P(y, z) \equiv \sum_{p,q} a_{p,q} y^q z^p = 0, \quad (y, z) \in \mathbb{C}^2, \quad (53)$$

where  $P$  is a polynomial of degree  $N$ . Actually, from eq. (32) it follows that  $P_X$  has degree  $N$  in  $y$ , but the  $a_i(z)$ 's could be any analytic functions on the punctured Riemann sphere. This means that they could be expressed by means of Laurent series in  $z$ . However in order to preserve the string interpretation we will limit ourselves to  $a_i(z)$ 's which are Laurent polynomials. Even more, in the following we will explicitly consider only  $a_i(z)$ 's which are polynomials in  $z$  in such a way that  $P(y, z)$  has overall degree  $N$ . This renders our discussion less general but far simpler.

The locus in  $\mathbb{C}^2$  of the solutions  $(y, z)$  of (53) is a *plane curve*. The independent non-vanishing coefficients  $a_{p,q}$  can be varied without changing, in general, the topological type  $(h, n)$  of the curve, where  $h$  is the genus and  $n$  is the number of punctures of the curve. They are the *moduli* of the plane curve. Counting them is necessary in order to see whether the moduli space of MST coincides with the moduli space of IIA superstring theory, or more realistically to what extent  $\mathcal{M}_N^{(h,n)}$  approximates  $\mathcal{M}^{(h,n)}$ .

In this section we discuss plane curves, in the next section their moduli.

The literature on plane curves, and, more generally, on algebraic curves is vast (see for instance [41, 42, 43]), and we will be using many well-known results. However one should bear in mind a peculiarity of our problem which is not usually considered in the textbooks on the subject: the question of punctures. So let us discuss first how to generate punctures. Later on we will see how to produce curves with non-zero genus. It will soon be clear that the generic plane curves from MST are in fact *singular*.

## 7.1 Punctures on plane curves

We recall that we interpret the Riemann surface defined by the relevant branched covering of the cylinder as the classical carrier of a string process. The basic information about branched coverings is of course determined by the branch points: this information is contained in the discriminant. The discriminant  $\delta$  of (32) is proportional to  $\Delta^2$ , where  $\Delta$  was defined in (37). The zeroes of the discriminant define the branch points and their multiplicity gives the multiplicity of the branch points, where the multiplicity or ramification index of a branch point is defined as the number of sheets which come together at that point, minus one; therefore branch points that involve only two sheets are called simple.

The branch points at  $z \neq 0, \infty$  represent joining and splitting processes of the string. Generically, when the branch point is simple, we have the joining of two strings to form a unique string or the splitting of one string into two. We may also have multiple branch points, in which more than two incoming or outgoing strings are involved. However the latter are limiting cases of the former and, from now on, as far as branch points at  $z \neq 0, \infty$  are concerned, the emphasis will be on simple branch points.

The inverse images under  $\pi$  of  $z = 0$ ,  $z = \infty$  are punctures in  $\Sigma$  with a definite string interpretation: they represent the points where incoming strings enter (outgoing strings leave) the process represented by the Riemann surface  $\Sigma$ .

It has to be kept in mind that, in MST, the counterimages of  $z = 0$  and  $z = \infty$  are distinguished points with an associated physical meaning. This is to be contrasted with the usual mathematical



treatment of branched coverings of  $\mathbb{CP}^1$ , where these points do not play any particular role. This remark will become extremely important below, in connection with the discussion about moduli space.

Let us discuss further properties of punctures corresponding to  $z = 0$  (an analogous discussion holds for  $z = \infty$ ). The counterimages of  $z = 0$  by  $\pi$  may be  $N$  distinct points, i.e. the solutions of the algebraic equation (32) at  $z = 0$  may be all distinct. In such a case we say we have  $N$  small incoming strings (of length 1 each). This is the case of the two incoming strings of the example considered in section 2. However, in general, the inverse image of  $z = 0$  may contain several branch points  $P_1^{in}, \dots, P_s^{in}$ , with multiplicity  $l_1 - 1, \dots, l_s - 1$ , respectively (if  $z = 0$  is a singular point of eq. (32) it has to be desingularized first, see below). In this case the process represented by  $\Sigma$  involves  $s$  incoming strings of length  $l_1, \dots, l_s$ , respectively. The physical interpretation of the string length has been given in [40, 13]. In the framework of the light-cone quantization of type IIA superstring, the string length is identified with the momentum component  $p^+ = p^9 + p^0$  of the string in suitably normalized units. Here 0,9 are of course the time and longitudinal direction of the ambient space, which do not explicitly appear in (21).

Let us see an example. Suppose  $y = y_1$  is a branch point of multiplicity  $l - 1$  in the counterimage of  $z = 0$ . This means that we have  $l$  roots of (32). For example,  $y^{(i)} \sim y_1 + \eta^i z^{1/l}$ ,  $i = 0, \dots, l - 1$  and  $\eta = \exp(2\pi i/l)$ . In other words  $l$  sheets of the covering join along a cut starting at  $y_1$ . The counterimage of a circle around  $z = 0$  in the  $z$ -plane contains a curve around  $y = y_1$  on the covering that closes after crossing the cut  $l$  times, i.e. we have an incoming string of length  $l$ . Therefore an easy rule to compute the length of an asymptotic string at a branch point in the inverse image of  $z = 0$  is to count the number of sheets that meet there. Alternatively such length can be seen as the period of the differential  $d \ln z$  around the point  $y = y_1$  of the covering. In fact  $y - y_1$  is a good coordinate near  $y_1$  and  $d \ln z = k d \ln(y - y_1)$ . The same conclusion can be drawn if the roots are like  $y^{(i)} \sim y_1 + \eta^i z^{j/l}$ , where  $j$  and  $l$  are relatively prime integers. A similar discussion can be carried out for the counterimages of  $z = \infty$  as well: for instance, in the  $N = 2$  example of section 2, the point at  $z = \infty$  is a simple branch point corresponding to a string of length 2.

Summarizing, punctures are the sites on the embedded Riemann surface, that is on the corresponding plane curve, where the incoming strings enter and the outgoing strings exit.<sup>2</sup> They are the counterimages by  $\pi$  of  $z = 0$  and  $z = \infty$ , respectively. If any such point on the plane curve is a branch point of multiplicity  $l - 1$ , then the corresponding incoming or outgoing string has length  $l$ . Incidentally, since eventually we want to take the large  $N$  limit, we are especially interested in the case when  $l$  is comparable with  $N$ . In the ordinary treatment of compact Riemann surfaces, if these points are regular, they are in no way special and must be considered on the same ground as all the other regular points (this can be seen for example by using projective coordinates). In our approach, on the contrary, this is not the case. As we have pointed out, the length of an incoming or outgoing string is interpreted as the  $+$  component of the momentum in the light-cone framework. Therefore the multiplicities of the branch points in the inverse image of  $z = 0, \infty$  have a precise physical meaning. Two processes that differ by these multiplicities must be kept distinct, even if, say, the topological type is the same.

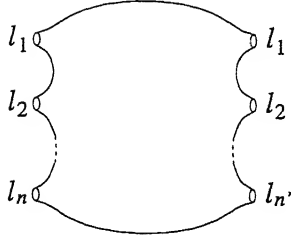
This is the picture of MST at strong coupling. At finite coupling  $g$  the string interpretation of instantons persists, but the dressing factor  $L$  has the effect of blurring it by smearing the string interactions.

## 7.2 Some examples

Before we continue the general discussion of plane curves, let us present some concrete examples of cases which are not unfamiliar in the physical literature.

We would like first to describe in detail how the genus zero (tree level) string interactions can be reproduced with a suitable form of the coefficients in the spectral equation (32) or (53). In

<sup>2</sup>In all the figures below we show the incoming and outgoing strings not as punctures, but as macroscopic strings in order to stress their different lengths.

Figure 3: Tree level process of  $n$  to  $n'$  strings.

the genus zero sector any Riemann surface  $\Sigma$  is a punctured sphere, realized as a  $N$ -fold branched cover of the  $z$ -sphere.

Assume we have  $n$  incoming and  $n'$  outgoing strings of lengths  $l_i$  and  $l'_j$ , ( $i = 1 \dots n, j = 1 \dots n'$ ), respectively (see the figure). From a physical point of view, we have seen that the length of a string is interpreted as the  $+$  component of its light cone momentum. We recall that the relation

$$\sum_i l_i = \sum_j l'_j = N, \quad (54)$$

must hold due to conservation of the momentum. We have also seen that the length of an incoming string  $i$  being  $l_i > 1$  means that the cover has a branch of order  $l_i - 1$  at  $z = 0$ , and likewise for outgoing strings at  $z = \infty$ .

Our aim here is to construct a polynomial  $P$  which underlies such a string process. Let us tackle this problem by studying the  $N$ -fold covering as a holomorphic projection from  $\Sigma$  to  $\mathbf{CP}^1$ . As we have already noticed, the coordinate  $z$  does represent such a projection as a meromorphic function on  $\Sigma$ : punctures manifest themselves as zeroes or poles of appropriate orders  $l_i, l'_j$ . The condition (54) means in this picture that the number of zeroes minus the number of poles, with multiplicity, is zero (this is the degree of the corresponding divisor).

Proceeding in this direction, we construct the generic meromorphic function in terms of a global coordinate on  $\Sigma$ , which we can take to be  $y$  itself. This is a useful simplification, which is not possible in higher genus cases.

The generic meromorphic function satisfying the above requirements on zeroes and poles is given by the following rational map:

$$z = K \frac{(y - y_1)^{l_1} (y - y_2)^{l_2} \dots (y - y_n)^{l_n}}{(y - y'_1)^{l'_1} (y - y'_2)^{l'_2} \dots (y - y'_{n'})^{l'_{n'}}}. \quad (55)$$

This map depends on  $n + n'$  parameters, in addition to the constant  $K$ : it fixes the  $n + n'$  punctures on  $\Sigma$  to be located at the points  $y_i$  and  $y'_j$ . The case of  $y_i$  or  $y'_j = \infty$  is a limiting case of the above formula when the relevant factor is absent. Let us verify that (55) gives the right behaviour at  $z = 0$  and  $z = \infty$ , see [19]. An example will suffice. Near  $y_1$  we can write  $z \sim (y - y_1)^{l_1}$ , therefore  $y \sim y_1 + z^{1/l_1}$ , which is exactly the behaviour considered above.

Now we can make a first exercise of moduli counting. Let us recall that the moduli space of the Riemann sphere with  $p$  punctures is  $p - 3$ . To count the moduli in (55), we first notice that we have  $n + n' + 1$  free parameters. Of these,  $K$  corresponds to a rescaling of the  $z$  coordinate; then we can use  $PSL(2, \mathbf{C})$  to reabsorb three parameters among the  $y_i, y'_j$ . As a result the meromorphic function describes spheres with  $n + n' - 3$  moduli, as expected.

Now, in order to see whether these curves are reproduced within MST, we try to cast (55) in the form (33). One sees immediately that (33) corresponds to curves where one of the outgoing punctures is at infinity, say  $y'_1 = \infty$ . Given that, the above map is indeed of the form of (33) with coefficients  $a_i$  which are at most linear in  $z$ :

$$y^N + a_{N-1}y^{N-1} + \dots + a_0 = 0, \quad a_i = \alpha_i z + \beta_i. \quad (56)$$

The generic polynomial of this form corresponds to a curve which has all  $l, l' = 1$ , i.e. it has no branches at the  $2N$  punctures, and depends on  $2N$  parameters. Of these, three can be ignored, since they correspond to transformations that leave  $y'_1 = \infty$ : a rescaling of  $z$ ; a shift of  $y$  and a rescaling of  $y$ . They are the remnant of  $PSL(2, \mathbb{C})$  which keeps  $y'_1 = \infty$ .

Therefore (33), or (53), contains the right  $2N - 3$  moduli of spheres with punctures.

The cases when some punctures are branched, are limiting cases of the previous curve when two or more punctures coincide. This can be easily seen from the meromorphic map (55). Therefore, for each  $l_i > 1$ , we have to enforce  $l_i - 1$  conditions on the parameters  $\alpha_i, \beta_i$  of the spectral equation. Thus the free parameters are, as expected:

$$2N - \sum_{i=1}^n (l_i - 1) - \sum_{j=1}^{n'} (l'_j - 1) - 3 = n + n' - 3.$$

We conclude that at genus zero MST reproduces, via (53), the full  $n + n' - 3$  moduli.

In the case of curves with non-vanishing genus, one would be tempted to proceed in the same way, that is to construct the meromorphic projection  $\Sigma \rightarrow \mathbb{CP}^1$  and then invert it. It is rather easy to construct the meromorphic function  $z$  corresponding to genus 1. However we come immediately across a novel feature which was absent in genus 0, but has dramatic consequences for the moduli counting.

The point is that the punctures on  $\Sigma$ , represented as zeroes and poles of the meromorphic function, cannot be arbitrary. This is a feature of the torus and of higher genus curves. There is a condition that they have to satisfy, which is the price we have to pay to be able to represent the punctured surface as an algebraic curve. Mathematically speaking, the divisor of a meromorphic function is not a generic divisor of degree zero, but is a principal one, which amounts to some extra condition on the punctures. The same condition was absent on the Riemann sphere because there every divisor of degree zero is principal.

To see which condition appears, let us represent explicitly the meromorphic function using a coordinate  $t$  taking values in the fundamental parallelogram. On a torus a meromorphic function can be represented as the ratio of products of translated theta functions:

$$z = K \frac{\theta(t - t_1)^{l_1} \theta(t - t_2)^{l_2} \cdots \theta(t - t_n)^{l_n}}{\theta(t - t'_1)^{l'_1} \theta(t - t'_2)^{l'_2} \cdots \theta(t - t'_{n'})^{l'_{n'}}}, \quad (57)$$

Now, for  $z$  to be single valued, the  $t_i, t'_j$  have to satisfy a condition, that is the vanishing of the Abel-Jacobi map:

$$\sum_i l_i t_i - \sum_j l'_j t'_j = 0 \mod \Gamma. \quad (58)$$

where  $\Gamma$  is the group of periods, which for the torus is the usual lattice of complex translations:  $\Sigma = \mathbb{C}/\Gamma$ .

It is instructive to look at the case of the propagator of a long string at genus one. In this case we have  $l_1 = l'_1 = N$ , while  $l_i = l'_i = 0$  for  $i \neq 1$ . We have the insertion on the torus of an incoming and an outgoing string of length  $N$ , at two points. By translation we can bring one of them at the origin and the other at, say,  $t'_1 = 0$  and  $t_1 = \bar{t}$ . The above condition is in this case:

$$N\bar{t} = 0 \mod \Gamma, \quad (59)$$

and we see that  $\bar{t}$  has to lie on the lattice  $\Gamma/N$  indicated in the figure. In this case of the torus with two punctures one expects in general two complex moduli. Here we see that we have one complex modulus  $\tau$  implicitly contained in the  $\theta$ -function, plus one discrete modulus  $\bar{t}$ . We can see from here that at finite  $N$  we have some limitations on the possible diagrams we can realize, however as  $N$  become large, the lattice  $\Gamma/N$  fills the plane and we recover the continuous modulus.

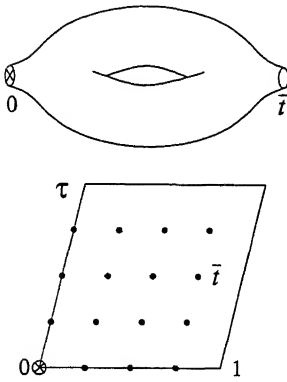


Figure 4: Discrete modulus

In the next section we will discuss in general the limitations of this kind. Therefore we leave this subject at this point and discuss other aspects concerning genus one curves.

The next thing we would like to do is to mimic the genus 0 case by inverting eq. (57). This is certainly possible locally, but, unlike the genus 0 case, we will not find in general a polynomial equation of the type (32). Therefore constructing the meromorphic projection (57) gives us only limited information about plane curves. In fact, what one expects is that the plane curve corresponding to (57) is in general in a singular representation (see below).

It is then necessary to study singular plane curves.

### 7.3 Plane curves and singular plane curves

At the beginning of this section we have called plane curves the locus of points which are solution of an equation like (53) in  $\mathbb{C}^2$ . This definition is too generic and lends itself to ambiguities. For example, we know the coordinates  $y$  and  $z$  are not on the same footing in MST. A  $z$  rescaling (at strong coupling) is a symmetry of any process in MST, but no other  $PSL(2, \mathbb{C})$  transformation on  $z$  is a symmetry transformation of a string process ( $z \rightarrow 1/z$  is a symmetry transformation of the theory, not of a single process). As for  $y$  it is not clear which coordinate transformations are a symmetry.

We resolve this and other ambiguities by embedding our curves in  $\mathbb{CP}^2$ : we introduce the homogeneous coordinates  $x_0, x_1, x_2$  with  $z = x_1/x_0, y = x_2/x_0$ . By multiplying (32) by a suitable power of  $x_0$  we obtain the equation of the curve in  $\mathbb{CP}^2$  in the form

$$F(x_0, x_1, x_2) = 0. \quad (60)$$

where  $F$  is the polynomial in  $x_0, x_1, x_2$  determined by  $P_X$ .

Then the coordinate transformations that do not change the curve are in general those of  $PGL(3, \mathbb{C})$ . However, as we said above, the points  $z = 0, \infty$  should be fixed in MST. This means that  $x_0 = 0$  and  $x_1 = 0$  should not be modified by any transformation. In conclusion the coordinate transformations that give rise to physically indistinguishable processes in MST, are those of the subgroup  $\mathcal{H} \subset PGL(3, \mathbb{C})$  defined by

$$\begin{pmatrix} x'_0 \\ x'_1 \\ x'_2 \end{pmatrix} = \begin{pmatrix} * & 0 & 0 \\ 0 & * & 0 \\ * & * & * \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ x_2 \end{pmatrix}. \quad (61)$$

In terms of  $y$  and  $z$ , these transformations include rescalings of  $y$  and  $z$  and linear transformations  $y \rightarrow y + \alpha z + \beta$ , with complex constants  $\alpha$  and  $\beta$ . They are acceptable coordinate transformations which involve 4 complex parameters. This fits our counting of the independent parameters in the previous subsection.

From now on, although we keep speaking mostly in terms of  $y$  and  $z$ , we always understand the corresponding formulation in terms of  $x_0, x_1, x_2$ . For example, a transformation like  $z \rightarrow 1/z$  must be accompanied by  $y \rightarrow y/z$  in order for us to remain within  $\mathbf{CP}^2$ . The latter is a compact space, therefore embedding the curves in it means compactifying them by filling the punctures with suitable points in  $\mathbf{CP}^2$ . These points will however always remain distinct due to the particular form of the symmetry subgroup  $\mathcal{H}$  we have chosen.

Given a curve defined by (60), the points in it where all partial first order derivatives vanish are *singular points*. When no singular point is present the curve is smooth. However this can happen only when its genus is  $\frac{1}{2}(d-1)(d-2)$ , where  $d$  is the degree of the curves, i.e. the degree of the polynomial  $P(y, z)$ : in our case  $d$  coincides with  $N$ . Therefore smooth plane curves cover only a very limited subset of the curves we need. One can lower the genus of the plane curve, while keeping the degree constant, by allowing for singularities. This means two important things: first, for finite  $N$  there exists an upper bound  $\frac{1}{2}(N-1)(N-2)$  on the genus of the Riemann surfaces which define the *core* of the stringy instantons; second, far from discarding singular curves, as one would be tempted to do as a first approach, we have to take them into account, they are bound to fill up most of the moduli space of plane curves. As we will see, singular curves are a happy occurrence, not a nuisance.

Singular curves play a major role in MST. For example, eq. (57) above, when written in homogeneous coordinates reveals a singularity corresponding to the point  $z = \infty$ . Singularities can be resolved so as to recover smooth Riemann surfaces (which are not anymore plane curves, in general). Given a singular curve  $\Sigma$  and the set  $\mathcal{S}$  of its singular points, a resolution of  $\Sigma$  is a smooth surface  $\tilde{\Sigma}$ , (usually embedded in a larger space than the original curve), together with a holomorphic projection  $\varpi : \tilde{\Sigma} \rightarrow \Sigma$ , such that its restriction  $\tilde{\varpi} : \tilde{\Sigma} - \varpi^{-1}(\mathcal{S}) \rightarrow \Sigma - \mathcal{S}$  is a biholomorphism.

In words, a resolution can be locally achieved by replacing a singular point by some space. A handy way is to replace the singular point by a sphere - called *exceptional divisor*. This is the well-known procedure of blowing up the singularity. A well-defined algorithm allows us to compute the genus of the desingularized curve, see [21].

A few examples of plane curves, including singular ones are given in Appendix B.

## 7.4 The role of singularities

It is impossible in general to represent Riemann surfaces by means of smooth plane curves embedded in the two complex dimensional space spanned by the coordinates  $y$  and  $z$ . One can say that singular plane curves within stringy instantons are the ordinary tools MST uses in order to reproduce the string interaction configurations required by string theory (actually, as will be seen in the next section, only in the  $N \rightarrow \infty$  limit is this completely true).

Far from representing a problem, singular plane curves are most welcome. They come with a gratifying bonus: the solution of a serious problem for the identification of MST at strong coupling with string theory. This identification is possible if string theory is formulated in the light-cone gauge. In MST, (21), ten dimensions enter into the game, two world-sheet dimensions plus eight transverse dimensions represented by the (diagonal)  $X^i$ . At first sight they seem to have a different nature, however it is clear that in a light-cone framework the two world-sheet dimensions are to be interpreted as representatives of the time and longitudinal dimensions, denoted 0 and 9, which bring the total of physical dimensions to ten. Now, stringy instantons characterized by a smooth plane curve, extend over four out of these ten dimensions. In other words it would seem that MST at strong coupling can only describe four-dimensional string processes. If this were true it would be hard to justify the correspondence MST — string theory.

However singular curves offer a solution to this problem. Singular curves become smooth if one enlarges the space where they are embedded. The standard way to resolve a singularity is to blow it up (see above), which means that a singular point is replaced by a two-dimensional sphere. For example, we have already pointed out that curves in  $\mathbf{CP}^2$  with nodes (a node is the simplest possible type of singularity) only, can be smoothed out by embedding them in  $\mathbf{CP}^3$ , i.e. by adding

two dimensions. It is natural to interpret this by saying that the corresponding string process extend over six (instead of four) dimensions. It is not difficult to imagine processes that extend over more (up to ten) dimensions. To better convince ourselves of this fact we can take the reverse point of view. Suppose we want to embed these higher (than four) dimensional processes within the instantons of the 2D field theory (21). The only possibility is to squeeze (project) them to the appropriate four dimensions: such operation of projecting gives rise to singularities. We suggest that the true significance of singular plane curves is given by their representing higher (than four) dimensional processes.

## 7.5 Summary

*In MST Riemann surfaces supporting string interactions make their appearance in the form of plane curves. Most of them are singular and need to be desingularized. The coefficients of the algebraic equations that define such plane curves are (up to symmetry identification) to be identified with moduli. The point is now to see to what extent the moduli of plane curves cover the moduli space of Riemann surfaces which are needed in string interaction theory.*

## 8 The moduli space of MST

Above we have seen that the genus of plane curves in MST is limited, at finite  $N$ , by an upper bound. There is another limitation to our capability to represent plane curves within MST which comes from the presence of punctures. In fact the presence of punctures on the Riemann surface entails the consequence that the *moduli space of plane curves of genus  $h$  with  $n$  punctures is a discretized version of the the moduli space of genus  $h$  Riemann surfaces with  $n$  punctures*, whose complex dimension is  $3g - 3 + n$ . A good parametrization of the moduli space fit for string interaction theory is provided by Mandelstam's variables, [40, 39]. By making a comparison with Mandelstam's parametrization, we will find out that  *$h$  of the Mandelstam complex parameters are actually discrete for the plane curves that appear in MST.*

The origin of such discretization can be briefly described as follows. The coordinate  $z$  we have introduced above, can be naturally regarded as a meromorphic function on a given plane curve (it is a realization of the projection  $\pi : \Sigma \rightarrow \mathbb{C}$ ). The counterimages of  $z = 0$  and  $z = \infty$  form a principal divisor in  $\Sigma$ . This entails, by Abel's theorem,  $h$  discretizing conditions on the parameters describing the plane curve. A detailed analysis shows that this imposes  $h$  of the Mandelstam parameters to be discrete.

*However when  $N \rightarrow \infty$  these discrete parameters become continuous and, in addition, the upper limit on the genus we mentioned above becomes ineffective.* Therefore for large  $N$  MST recovers the full moduli space of string theory. We recall that, for finite  $N$ , also the  $p^+$  components of the momenta of the incoming outgoing strings are discrete, so that again continuity is recovered only for  $N \rightarrow \infty$ . Therefore, *a complete description of string interaction theory can be truly achieved by MST only in the large  $N$  limit.* It is nevertheless remarkable that genus 0 processes (with discrete  $p^+$  components of the external momenta) are exactly described by MST also for finite  $N$ .

A very convenient way to proceed is to make a comparison with the Mandelstam parametrization of the moduli space of Riemann surfaces with punctures [40]. To this end, let us first review some basic facts about the realization of Mandelstam diagrams. We refer to [39] for a complete account of the following very quick review, after which, we will examine the consequence of the main new input from MST, that is holomorphicity of the covering map which defines the Mandelstam diagram. The result will be a set of constraints on the kinematical data of the diagram which turn out to be a quantization condition for some of the Mandelstam parameters. In the large  $N$  limit these constraints loosen their effectiveness and allow us to recover the full moduli space of the string diagrams.

Let  $\Sigma$  be a compact Riemann surface of genus  $h$  and let  $\omega_I$ ,  $I = 1, \dots, h$  be a set of holomorphic differentials on  $\Sigma$  normalized by  $\oint_{\alpha_J} \omega_I = \delta_{IJ}$ , while  $\oint_{\beta_J} \omega_I = \Omega_{IJ}$  is the period matrix. We fix  $n$

punctures  $\{Q_1, \dots, Q_n\}$  on  $\Sigma$  and define the divisor  $D = Q_1 \cdot \dots \cdot Q_n$ . We also introduce a set of  $n$  real numbers  $R = \{r_1, \dots, r_n\}$  such that  $\sum_i r_i = 0$ .

Now, let  $\omega$  be the differential which is holomorphic on  $\Sigma \setminus D$  with simple poles at  $D$  with  $\text{res}_{Q_i} \omega = r_i$  and  $\text{Re} \oint_{\alpha_i} \omega = 0 = \text{Re} \oint_{\beta_i} \omega$ .

In [39] it was shown how the above differential defines a nice procedure which allows us to look at  $\Sigma$  as a *topological* covering of a cylinder: one can easily decompose  $\Sigma$  into pants along the level lines of the function  $\tau(P) \equiv \text{Re} \int^P \omega$ . In this sense,  $\omega$  induces on  $\Sigma$  the structure of a Mandelstam diagram. The Mandelstam parameters are the twist-angles  $\theta_b$ ,  $b = 1, \dots, 3h + n - 3$ , along the junctures of the pants decomposition and the relative time coordinates  $\tau_a - \tau_0$ ,  $a = 1, \dots, 2h + n - 3$ , of the  $2h + n - 2$  interaction points.  $h$  additional real parameters are the internal light-cone momenta  $p_I^+ = \oint_{\alpha_I} \omega$ . Altogether they form a set of  $6h - 6 + 2n$  real parameters. In [39] it was shown that these parameters represent good coordinates on the moduli space of genus  $h$  Riemann surfaces with  $n$  punctures,  $\mathcal{M}_{h,n}$ .

To complete the picture we identify the set  $R$  with the  $+$  components of the external light-cone momenta of the diagram, i.e. the periods of  $\omega$  around the punctures. We also have the relations  $\oint_{\beta_i} \omega = \frac{i}{2\pi} p_K^+ \mathcal{W}_{Kb}^I \theta_b$ , where  $\mathcal{W}^I$  are integer-valued matrices which depends on the pants decomposition of the Riemann surface and its intersections with the  $\alpha$  and  $\beta$  cycles.

Our strategy now is the following. We first construct an explicit form for  $\omega$ , in terms of the prime-form, the  $\omega_I$ 's and the period matrix of  $\Sigma$ . Then we compare this  $\omega$  with the one that comes from MST. The relevant new input consists in the fact that MST induces on  $\Sigma$  the structure of a *holomorphic* covering of the Riemann sphere (as usual we consider the latter instead of the cylinder). By this we mean that, if  $z : \Sigma \rightarrow \mathbb{CP}^1$  is the covering map in the MST scheme, the coordinate  $z$  is a meromorphic function on  $\Sigma$ . The role of  $\omega$  in MST is played by  $d \ln z$ , therefore we have to identify them. This condition becomes a constraint on the data of the Mandelstam diagram. In fact, it means that  $D^R \equiv Q_1^{r_1} \cdot \dots \cdot Q_n^{r_n}$ , being the divisor of the meromorphic function  $z$ , is a principal divisor on  $\Sigma$ , in particular  $r_i \in \mathbb{Z}$ . As a consequence, some constraints appear in the data of the Mandelstam diagram and these conditions induce a complex codimension  $h$  slicing of the moduli space. This can be seen as follows.

Let  $\omega_{P_+P_-}$  be the holomorphic differential on  $\Sigma \setminus \{P_+, P_-\}$  with simple poles at  $P_{\pm}$  with residues  $\pm 1$  and imaginary periods. It can be written as

$$\omega_{P_+P_-}(P) = d_{(P)} \ln \left[ \frac{E(P, P_+)}{E(P, P_-)} \cdot e^{2\pi i \text{Im} \int_{P_-}^{P_+} \omega_I \Omega^{(2)}{}_{IJ}^{-1} \int^P \omega_J} \right] = d_{(P)} \ln H(P, P_+, P_-), \quad (62)$$

where  $E(P, Q)$  is the prime form on  $\Sigma$ ,  $\Omega^{(2)}$  is the imaginary part of the period matrix and  $d_{(P)} = dP \cdot \frac{\partial}{\partial \bar{P}}$ .

In terms of the above differentials we can write

$$\omega = \sum_{l=1}^{n-1} k_l \omega_{Q_l Q_{l+1}}, \quad (63)$$

where  $k_i - k_{i-1} = r_i$  and  $k_0 = 0 = k_n$ ; substituting (62) into (63) we obtain

$$\omega(P) = d_{(P)} \ln \tilde{z}(P)$$

where

$$\tilde{z}(P) = \prod_{l=1}^{n-1} [H(P, Q_l, Q_{l+1})]^{k_l}. \quad (64)$$

Now, as anticipated above, we make the identification  $\omega = d \ln z$ . This requires that  $\tilde{z} = z$  up to a multiplicative constant, which implies that  $\tilde{z}$  is a well defined meromorphic function on  $\Sigma$ . On the one hand this imposes that the residues  $r_i$  be quantized in integer values. On the other

hand it requires that the differential  $d\bar{z}$  have vanishing periods along  $\alpha$  and  $\beta$  cycles. The latter condition is fulfilled iff

$$\sum_{l=1}^n r_l \int^{Q_l} \omega_I = m_I + n_J \Omega_{JJ} \quad (65)$$

for some  $m_I, n_I \in \mathbb{Z}$ . At this point the situation is clear: (65) is the vanishing condition for the Abel map and says that the divisor  $D^R$  is principal.

Conversely, let  $z$  be a meromorphic function on  $\Sigma$  and  $D^R$  its divisor. By definition (65) holds and  $\text{res}_{Q_l} d_{(P)} \ln z = r_l$ .

Notice that the periods of  $\omega$  are quantized in integral values as

$$\oint_{\alpha_I} \omega = 2\pi i n_I \quad \text{and} \quad \oint_{\beta_I} \omega = -2\pi i m_I, \quad (66)$$

and this condition is equivalent to (65).

Eq. (66) means that the internal light-cone momenta of the diagram are quantized and that, in addition, there are  $h$  discretizing constraints on the twist-angles of the Mandelstam diagram. Since these variables, together with the relative interaction times which have been left untouched, are the coordinates of the moduli space, we are left with a discrete slicing of the moduli space  $\mathcal{M}_{h,n}$ , each slice being of complex dimension  $2g - 3 + n$ . This discretized moduli space is what we have called  $\mathcal{M}_N^{(h,n)}$  in section 5.4.

One can verify that in all the genus one examples we have considered in section 7, the counting of independent parameters matches the formula  $2h + n - 3$ . We believe that, for any topological type  $(h, n)$ , one can construct plane curves with  $2h + n - 3$  independent parameters.

A confirmation of this result comes from an estimate of the moduli space of stringy instantons. Since in the  $Y$  factor there are no free parameters, the moduli space of stringy instantons must coincide with the free parameters contained in  $M$ , i.e. with the moduli space of plane curves. The estimate carried out in [21] confirms the above evaluation of the continuous dimension of the moduli space of the latter.

In the large  $N$  limit, however, the quantization condition disappears in a continuum of values.

$$\lim_{N \rightarrow \infty} \frac{1}{N} [Z^h \oplus \Omega Z^h] = \mathbb{C}^h. \quad (67)$$

Simultaneously, for large  $N$  also the bound  $\frac{1}{2}(N-1)(N-2)$  on the genus of the plane curves in MST, becomes ineffective. It is therefore sensible to argue that for large  $N$  one recovers the full moduli space of string theory.

## 8.1 Summary

*The moduli space of Riemannian instantons which appear in MST, is only an approximated version of the moduli space of Riemann surfaces which appear in string interaction theory. In particular  $h$  among the former are a discrete version of  $h$  among the latter. It is however reasonable to assume that in the large  $N$  limit the two spaces tend to coincide.*

## 9 Comments

We have seen that in MST Riemann surfaces are generated as classical solutions of the equations of motion. More precisely they come dressed by a factor that tends to 1 in the strong coupling limit. Therefore, in this limit, we are left with pure Riemann surfaces with punctures, which can be thought of as carriers of a string interactions. We have seen that this leads to a consistent picture: the strong coupling limit action of MST is the Green-Schwarz action of IIA superstring theory plus a free Maxwell action; the latter guarantees that the path integral for a string interaction process



in strong coupling MST has the correct form; when  $N \rightarrow \infty$ , the amplitudes computed in string interaction theory and in strong coupling MST tend to coincide. It is therefore legitimate to claim that the strong coupling MST represents type IIA superstring theory. Recently new results have been found in this field, [45, 22]. In particular in [22] it has been shown that the strong coupling limit of Heterotic Matrix String Theory (a variant of MST with gauge group  $SO(N)$ ) describes the heterotic superstring theory.

MST is therefore a remarkable case of a Yang–Mills theory with a definite string interpretation. It is however clear from section 2 that this is not the only interesting case. Riemannian instantons exist for example also in a 4d Yang–Mills theory. They lend themselves to a string interpretation of (some limit of) Yang–Mills theory. In any case they represent a stimulating possibility which has not been exploited so far.

## 10 Appendices

### 10.1 Appendix A. Analysis of the sinh–Gordon equation

We discuss here an (approximate) analytic approach to the sinh–Gordon equation (14) with boundary conditions (13) and vanishing at  $z = 0, \infty$ .

Recalling that

$$\frac{\partial \zeta}{\partial z} = \sqrt{2g} \frac{\sqrt{z - z_0}}{z},$$

the approximate expression of  $\zeta$  in terms of  $z$  is

$$\begin{aligned} \zeta &\sim \frac{2\sqrt{2}g}{3z_0}(z - z_0)^{3/2}, \quad \text{for } z \sim z_0 \\ \zeta &\sim \frac{\sqrt{2}g}{2}\sqrt{z}, \quad \text{for } |z| \gg |z_0| \\ \zeta &\sim \sqrt{2}g\sqrt{-z_0} \ln z, \quad \text{for } |z| \ll |z_0|. \end{aligned}$$

If these were the exact expressions for  $\zeta$ , we could consider spherically symmetric solutions of (14), i.e. solutions depending only on  $r = |\zeta|$ . For them eq.(14) takes the form

$$\partial_r^2 u + \frac{1}{r} \partial_r u = 4 \sinh u \quad (68)$$

This is a form of the Painlevé III equation. The general form of the solutions of this equation are known, see [34] and references therein. Let us select the class of solutions with the following asymptotic behaviour:

$$u(r) \sim \alpha \ln r + 1, \quad r \rightarrow 0, \quad |\alpha| < 2 \quad (69)$$

$$u(r) \sim \gamma r^{-1/2} e^{-2r}, \quad r \rightarrow \infty. \quad (70)$$

The constants  $\beta$  and  $\gamma$  must be fine-tuned to  $\alpha$  in order to give rise to smooth solutions. However, in our case, we are not interested in the actual value of  $\beta$  and  $\gamma$ , therefore we can always adjust the parameters in such a way as to have a smooth solution. As for the bound  $|\alpha| < 2$ , in our case

$$u \sim -\frac{1}{3} \ln r$$

therefore the bound is satisfied.

Let us study now the properties of the solution. In the various regions we have the following asymptotic expressions:

$$r \sim \frac{2\sqrt{2}g}{3|z_0|} |z - z_0|^{3/2}, \quad \text{for } z \sim z_0$$

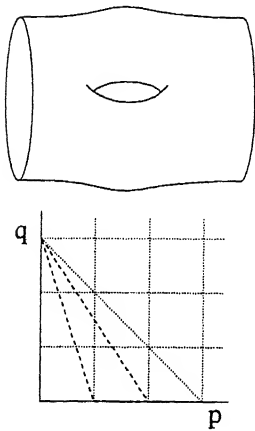


Figure 5: Self-energy of a string.

$$\begin{aligned} r &\sim \frac{\sqrt{2}g}{2} \sqrt{|z|}, \quad \text{for } |z| \gg |z_0| \\ r &\sim \sqrt{2}g \sqrt{|z_0|} \ln |z|, \quad \text{for } |z| \ll |z_0|. \end{aligned} \quad (71)$$

From this we see that, when, at fixed finite  $g$ ,  $z$  is near the origin and far to infinity, the solution tends to zero. The convergence to zero is more rapid the larger  $g$  is, the slope in  $g$  being of negative exponential type. Looking now at the first equation (71), we see that, even if we sit near  $z_0$ , we may still fall in the regime ( $r$  large) in which the solution is extremely small, provided  $g$  is large enough. In other words, for large  $g$  the solution shrinks around  $z_0$ , and, in the  $g \rightarrow \infty$  limit it becomes spike-like with support at  $z = z_0$ . We can say that, if we exclude a neighborhood of  $z_0$  of size proportional to  $(g)^{-2/3}$ , the solution decreases to zero more rapidly than any power of  $1/g$ .

We recall that the spherically symmetric solution is not the exact solution, but only an approximate one. However we expect the general behaviour of the true solution to be essentially similar, i.e. that it shrinks very rapidly around  $z_0$  as  $g \rightarrow \infty$ .

One can easily extend the previous analysis to the case in which  $a$  contains several distinct zeroes. Simply find suitable approximate expressions for  $\zeta$  near the zeroes of  $a$  and apply the previous approximate analysis. The conclusion will be that the solution shrinks very quickly around the branch points as  $g \rightarrow \infty$ .

## 10.2 Appendix B. Smooth and singular plane curves

This Appendix is devoted to some explicit examples of smooth and singular plane curves.

A useful tool in studying plane curves is the *Newton Polygon*. Let us consider the polynomial  $P(y, z)$  in (53). We associate to each monomial  $z^\alpha y^\beta$  in it a point  $p = \alpha$ ,  $q = \beta$  in a  $p, q$  plane. We obtain a set of points called the *carrier*: its convex hull is by definition the *Newton polygon* associated to the curve. From the Newton polygon one can deduce a lot of information concerning the curve. For the curves we consider the Newton polygon always contains the point  $(p = 0, q = N)$  and is contained in the equilateral triangle formed by the  $p$  and  $q$ -axis and by the line  $p + q = N$ .

We start with the case  $N = 3, g = 1$ , for which there is already a good variety of examples. These have the advantage that one can check the results by explicitly solving the cubic algebraic equation by means of Cardano's formula. We do not write down the algebraic equations, but simply the corresponding polygons. In the following figures the Newton polygon is the one delimited by the dashed lines. The coefficient of the monomials within or on the border of the Newton polygon are understood to be generic, unless otherwise specified.

The simplest process one can imagine is the string self-energy. This means that we have to look for a totally branched curve over  $z = 0$  and  $z = \infty$ . Remember that the polynomials giving

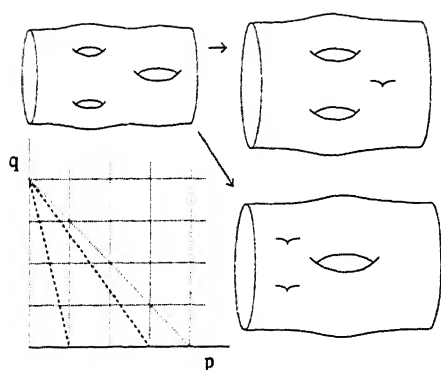


Figure 6: Shrinking cycles: totally branched quartics at genus two and one.

the solutions over these points are given by the points of the carrier on the  $q = 0$  and  $p + q = N$  lines respectively. So one simple solution is given by the carrier shown in figure 5; the generic case will be non-singular also at finite  $z$  and so the genus will be one. The presence of the points  $(1, 0)$  and  $(2, 0)$  ensures the nonsingularity of  $0$  and  $\infty$ ; the local behaviour around them is given by the upper side of the Newton polygon.

For other smooth examples, see [21]. Let us consider now examples of singular curves. Given an algebraic equation, a singularity may appear for some particular choice of the parameters. In this case one has simply to replace the finite hole in these figures by a hole shrunk to a point (for example see fig. 6); the curve becomes genus zero, i.e. a sphere with two identified points. This singularity is the simplest one, it is characterized by a non-vanishing Hessian and is called a *node*. All nodes can be viewed as two points identified: blowing up a node amounts to separating the points. For instance, consider the previous case (figure 5): the polynomial which corresponds to the diagram can be written as

$$P = y^3 + czy + z(z - a). \quad (72)$$

Imposing that a point be singular, one finds that a necessary (but not sufficient) condition is that its discriminant,  $\delta = z^2[27(z - a)^2 - 4c^3z]$ , have a multiple root. The double root at  $z = 0$  just signals that this point is another triple branch point, as we already know; imposing that the remaining factor be a square, one finds several values, of which for instance  $c = 0$  gives a triple branch at  $z = a$  and no singular point, and  $c = -3a^{1/3}$  gives instead a node.

This introduces us to our next task: to show how it is possible to describe low-genus highly branched curves. We will describe in detail the self-energy case. We take  $N = 4$ ; since we want total branching, we can choose a diagram like that in figure 6. The corresponding polynomial has the coefficients corresponding to the vertices of the polygon, and can also have coefficients corresponding to the points on the sides or in the interior (by the way the latter are always  $(N - 1)(N - 2)/2$  in number, if there is no singular point at  $0$  and at  $\infty$ , and count the genus of the corresponding smooth curve). Now we can look for singular cases in this family along the lines of the previous example; since already in this case computations become complicated, we restrict ourselves to the biquadratic case. In other words the polynomial we start with is

$$P = y^4 + bzy^2 + z(d + ez + fz^2); \quad (73)$$

its discriminant is

$$\delta = 16z^3(d + ez + fz^2)(4d + 4ez + 4fz^2 - b^2z)^2. \quad (74)$$

As before, the term  $z^3$  shows that the branching at  $z = 0$  is of order three, i.e. four sheets meet there. The other two terms mean the following. Solutions of a biquadratic equation are in general  $\pm y_{1,2}$ . Its discriminant can vanish in two cases: if  $y_1 = y_2$  or  $y_1 = -y_2$  — this is determined by the third term in (74) — in which case, at the corresponding value of  $z$ , there is a couple of double branch points; if  $y_1 = 0$  or  $y_2 = 0$ , which is determined by the second term, there is a single node.

If we choose the coefficients so that the third term is a fourth power, we have two nodes, and so genus one; if, instead, the coefficients are chosen so that the second term is a square, we have a single node, and so genus two. The situation is shown in figure 6.

### Acknowledgments

This review is based on joint work done with G.Bonelli, F.Nesti and A.Tomasiello, whom I would like to thank for their collaboration. This work was partially supported by EC TMR Programme, grant FMRX-CT96-0012, and by the Italian MURST for the program “Fisica Teorica delle Interazioni Fondamentali”.

### References

- [1] G. 't Hooft, *A Planar Diagram Theory for Strong Interactions*, Nucl.Phys. **B72** (1974) 461; *A Two-Dimensional Model for Mesons*, Nucl.Phys. **B75** (1974) 461.
- [2] see for example: D.J.Gross and W.Taylor, *Two-dimensional QCD and strings*, [HEP-TH9311072].
- [3] see for example: E.Brezin and S.Wadia *The large N expansion in quantum field theory and statistical physics: from spin systems to 2-dimensional gravity*, World Scientific, Singapore, 1993.
- [4] J.Maldacena, *The large N limit of superconformal field theories and supergravity*, Adv.Theor.Math.Phys. **2** (1998) 231, [HEP-TH9711200].
- [5] A.Polyakov, *The Wall of the Cave*, [HEP-TH9809057]
- [6] I.R.Klebanov and A.A.Tseytlin, *D-branes and dual Gauge Theories in Type 0 Strings*, [HEP-TH9811035].
- [7] M.B. Green, J.H. Schwarz, E. Witten, *Superstring Theory*, Cambridge Univ. Press, Cambridge 1987.
- [8] J. Polchinski, *String Theory*, Cambridge Univ. Press, Cambridge 1998.
- [9] T. Banks, W. Fischler, S.H. Shenker and L. Susskind, *M Theory As A Matrix Model: A Conjecture*, Phys.Rev.D **55** (1997) 5112 [HEP-TH9610043].
- [10] L. Motl, *Proposals on Nonperturbative Superstring Interactions*, [HEP-TH9701025].
- [11] T. Banks and N. Seiberg, *Strings from Matrices*, Nucl.Phys. B **497** (1997) 41 [HEP-TH9702187].
- [12] W. Taylor, *D-brane Field Theory on Compact Spaces*, Phys.Lett. **B394** (1997) 283 [HEP-TH9611042].
- [13] R. Dijkgraaf, E. Verlinde, H. Verlinde, *Matrix String Theory*, Nucl.Phys. **B500** (1997) 43 [HEP-TH9703030].
- [14] R. Dijkgraaf, G. Moore, E. Verlinde, H. Verlinde, *Elliptic Genera of Symmetric Products and Second Quantized Strings*, Comm.Math.Phys. **185** (1997) 197 [HEP-TH9608096].
- [15] H. Verlinde, *A Matrix String Interpretation of the Large N Loop Equation*, [HEP-TH9705029].
- [16] L. Bonora, C.S. Chu, *On the String Interpretation of M(atrix) Theory*, Phys.Lett. **B410** (1997) 142 [HEP-TH9705137].

- [17] T. Wynter, *Gauge fields and interactions in matrix string theory* Phys.Lett. **B415** (1997) 349 [HEP-TH9709029].
- [18] S.B. Giddings, F. Hacquebord, H. Verlinde, *High Energy Scattering of D-pair Creation in Matrix String Theory* Nucl.Phys. **B537** (1999) 260 [HEP-TH9804121].
- [19] G. Bonelli, L. Bonora and F. Nesti, *Matrix string theory, 2D instantons and affine Toda field theory*, Phys.Lett. **B435** (1998) 303 [HEP-TH9805071].
- [20] G. Bonelli, L. Bonora and F. Nesti, *String Interactions from Matrix String Theory*, Nucl.Phys. **B538** (1999) 100 [HEP-TH9807232].
- [21] G. Bonelli, L. Bonora, F. Nesti and A.Tomasiello, *Matrix String Theory and its Moduli space* [HEP-TH9901093], to be published in Nucl.Phys.B.
- [22] G. Bonelli, L. Bonora, F. Nesti and A.Tomasiello, *Heterotic Matrix String Theory and Riemann Surfaces* [HEP-TH9905092]
- [23] L. Susskind, *Another Conjecture about M(atrrix) Theory*, [HEP-TH9704080].
- [24] A. Bilal, *M(atrrix) theory: a pedagogical introduction*, [HEP-TH9710136].
- [25] T. Banks, *Matrix Theory*, [HEP-TH9710231].
- [26] D. Bigatti and L. Susskind, *Review of Matrix Theory*, [HEP-TH9712072].
- [27] Washington Taylor IV, *Lectures on D-branes, Gauge Theory and M(atrices)*, [HEP-TH9801182].
- [28] N.J. Hitchin, *The self-duality equations on a Riemann surface*, Proc. London Math. Soc. 55 (1987) 59; *Lie groups and Teichmüller space*, Topology **31** (1992) 449.
- [29] N. Hitchin, *Stable bundles and integrable systems*, Duke Math. Jour. **54** (1987) 91.
- [30] E. Markman, *Spectral curves and integrable systems*, Comp. Math. **94** (1994) 255.
- [31] C.T. Simpson, *Harmonic bundles on noncompact curves*, Jour.Am.Math.Soc. **3** (1990) 713.
- [32] R. Donagi and E. Witten, *Supersymmetric Yang-Mills theory and integrable systems*, Nucl.Phys. **B460** (1996) 299 [HEP-TH9510101].
- [33] M. Bochicchio *The large- $N$  limit of QCD and the collective field of the Hitchin fibration*, JHEP **01** (1999) 006 [HEP-TH9810015].
- [34] A.R. Its and V.Yu. Novokshenov, *The isomonodromic deformation method in the theory of Painlevé equations*, Lect. Notes Math. 1191, Springer-Verlag 1986.
- [35] E. Gava, J.F. Morales, K.S. Narain, G. Thompson, *Bound States of Type I D-Strings* [HEP-TH9801128]
- [36] E. Witten, *On S-duality in abelian gauge theory* [hep-th/9505186].
- [37] L.L. Ahlfors, *Open Riemann surfaces and extremal problems on compact subregions*, Comm. Math. Helv. **24** (1950) 100.  
J.D. Fay, *Theta functions on Riemann surfaces*, Lect.Not.Math. Vol.352, Springer-Verlag, Berlin 1973.
- [38] J.-L. Gervais and B. Sakita, *Extended particles in quantum field theories*, Phys.Rev.D **11**(1975)2943.

- [39] S.B. Giddings and S.A. Wolpert, *A triangulation of moduli space from light cone string theory* Comm.Math.Phys.**109** (1987) 177.
- [40] S. Mandelstam, *Dual resonance models*, Phys.Rep. **13** (1974) 259.
- [41] E. Brieskorn and H. Knörrer, *Plane Algebraic Curves*, Birkhäuser Verlag, Basel 1986.
- [42] P. Griffiths and J. Harris, *Principles of Algebraic Geometry* New York 1978.
- [43] F. Kirwan, *Complex algebraic curves* Cambridge 1992.
- [44] I.M.Gelfand, M.M.Kapranov, A.V.Zelevinsky, *Discriminants, resultants, and multidimensional determinants*, Boston, Birkhauser, 1994.
- [45] T.Wynter, *High energy scattering amplitudes in matrix string theory*, [HEP-TH9905087].



## Part E : QFT In $2 + 1$ Dimensions

26. Fractional Statistics And Chern-Simons Field Theory In  $2 + 1$  Dimensions

by Avinash Khare

27. Chern Simons Field And Composite Bosons In The Quantum Hall System

by R.Rajaraman





# 26. Fractional Statistics and Chern-Simons Field

Avinash Khare \*

Institute of Physics, Sachivalaya Marg, Bhubaneswar 751005, India

## Abstract

The question of anyons and fractional statistics in field theories in 2+1 dimensions with Chern-Simons (CS) term is discussed in some detail. Arguments are spelled out as to why fractional statistics is only possible in two space dimensions. This phenomenon is most naturally discussed within the framework of field theories with CS term, hence as a prelude to this discussion I first discuss the various properties of the CS term. In particular its role as a gauge field mass term is emphasized. In the presence of the CS term, anyons can appear in two different ways i.e. either as soliton of the corresponding field theory or as a fundamental quanta carrying fractional statistics and both approaches are elaborated in some detail.

## 1 Introduction

Many of us have wondered some time or the other if one can have nontrivial science and technology in two space dimensions; but the usual feeling is that two space dimensions do not offer enough scope for it. This question, to the best of my knowledge, was first addressed in 1884 by E.A. Abbot in his satirical novel *Flatland* [1]. The first serious book on this topic appeared in 1907 entitled *An episode of Flatland* [2]. In this book C.H. Hinton offered glimpses of the possible science and technology in the flatland. A nice summary of these two books appeared as a chapter entitled *Flatland* in a book in 1969 edited by Martin Gardner [3]. Inspired by this summary, in 1979 A.K. Dewdney [4] published a book which contains several laws of physics, chemistry, astronomy and biology in the flatland. However, all these people missed one important case where physical laws are much more complex, nontrivial and hence interesting in the flatland than in our three dimensional world. I am referring here to the case of quantum statistics. In last two decades it has been realized that whereas in three and higher space dimensions all particles must either be bosons or fermions (i.e. they must have spin of  $n\hbar$  or  $(2n+1)\hbar/2$  with  $n=0,1,2,\dots$  and must obey Bose-Einstein or Fermi-Dirac statistics respectively), in two space dimensions the particles can have any fractional spin and can satisfy *any* fractional statistics which is interpolating between the two. The particles obeying such statistics are generically called as *anyons* [5]. In other words, if one takes one anyon slowly around the other then in general the phase acquired is  $\exp(\pm i\theta)$ . If  $\theta=0$  or  $\pi$  (modulo  $2\pi$ ) then the particles are bosons or fermions respectively while if  $0 < \theta < \pi$  then the particles are termed as anyons.

From our experience with fermions and bosons it is well known that the question of spin and statistics can be properly handled only within the formalism of relativistic quantum field theory. Thus it is of interest to enquire if one can also understand the ideas of anyons and fractional statistics within the formalism of relativistic quantum field theory. This is the issue that we would like to discuss in this article.

Before I go into the details, one might wonder if our discussion is merely of academic interest? The answer to the question is no. In fact it is a surprising fact that two, one and even zero dimensional experimental physics is possible in our three-dimensional world. This is because of the third law of thermodynamics, which states that all the degrees of freedom freeze out in the limit of zero temperature, it is possible to strictly confine the electrons to surfaces, or even to

---

\*Email:khare@iopb.res.in

lines or points. Thus it may happen that in a strongly confining potential, or at sufficiently low temperatures, the excitation energy in one or more directions may be much higher than the average thermal energy of the particles, so that those dimensions are effectively frozen out. Of course, even then, at the basic level, the fundamental particles are certainly fermions or bosons. However, the most direct and appropriate discussion of the low energy behavior of a material is usually in terms of the quasi-particles. The hope is that at least in some of these cases the quasi-particles could be anyons. This hope has in fact been realized in the case of the fractionally quantized Hall effect where the quasi-particles are believed to be charged vortices i.e. charged anyons [6]. Recent experiments [7] seem to confirm the existence of fractionally charged excitations and hence indirectly of anyons.

The plan of the article is the following. In Sec.II, I first spell out as to why fractional statistics is only possible in two space dimensions. It turns out that the phenomenon of fractional statistics is most naturally discussed within the framework of field theories with CS term. As a prelude to this discussion, in Sec.III, I discuss the various properties of the CS term. In particular its role as a gauge field mass term and its behavior under the discrete transformations of parity (P) and time-reversal (T) is emphasized. In the presence of the CS term, anyons can appear in two different ways (i.e. either as soliton of the corresponding field theory or as fundamental quanta carrying fractional statistics) and both approaches are elaborated in some detail in the next three sections. The charged vortex solutions in Abelian Higgs model with CS term are obtained in Sec.IV, and it is pointed out that these charged vortices represent the first relativistic model for (extended) charged anyons. I also construct the charged vortex solutions in pure CS theory in both the relativistic and the non-relativistic settings. In Sec.V, I discuss an example of neutral relativistic anyons by considering the soliton solutions in the  $CP^1$  model with the Hopf term which is one of the *avatars* of the CS term. Finally, in Sec.VI, I elaborate upon the other approach in which fundamental fields of theories with CS term themselves carry fractional spin and obey fractional statistics.

## 2 Why Anyons in Only Two Dimensions?

Before we come to the question of fractional statistics, it might be worthwhile to understand as to why unlike in three and higher space dimensions, the eigenvalue of the spin angular momentum operator can take any fractional value in units of  $\hbar$ . The point is that the spin in two dimensions differs fundamentally from the spin in higher dimensions. This is because whereas in three and higher space dimensions, the spin angular momentum algebra is non-commutative i.e.

$$[S_i, S_j] = i\hbar \varepsilon_{ijk} S_k ; \quad i, j, k = 1, 2, 3 \quad (1)$$

in two space dimensions, it is a trivial commutative algebra since only one generator (say  $S_3$ ) is available which obviously commutes with itself. As a result, there is no analogue of the quantization of the angular momentum, which arises in three and higher space dimensions from the nonlinear commutation relation (1). Here  $\varepsilon_{ijk}$  is the completely antisymmetric tensor.

Now, in relativistic quantum field theory, there is a deep and profound connection between the spin and the statistics i.e. particles with half integer spin are fermions, satisfying Fermi-Dirac statistics, while those with integer spin are bosons, satisfying Bose-Einstein statistics. This immediately suggests that in two dimensions the particles may exhibit fractional (i.e. any) statistics. In a remarkable paper Leinaas and Myrheim [8] showed that this is indeed so. Before we come to a proper discussion about the statistics, it is worth clarifying as to what exactly one means by quantum statistics. In most text books on statistical mechanics, the term “quantum statistics” refers to the phase picked up by a wave function when two identical particles are interchanged, i.e, under the permutation of the particles. But this is slightly misleading and has been correctly criticized in the literature [9]. If the particles are *strictly identical*, the word permutation has no physical meaning since a given configuration and the one obtained by the permutation of the particle coordinates are merely two different ways of describing the *same* particle configuration. The term quantum statistics actually refers to the phase that arises when two particles are adiabatically transported giving rise to the exchange. In this book, we shall be concentrating on this definition of quantum

statistics. It is a coincidence that in three and higher dimensions, the two definitions, based on the permutation and the adiabatic exchange of two particles, coincide, but in two dimensions the two definitions give very different answers.

The key reason for the fractional statistics in two dimensions is the principle of indistinguishability of identical particles. It is one of the most important characteristics of quantum mechanics (vis a vis classical mechanics) and it has profound physical consequences. The principle is in fact older than quantum mechanics. It was introduced by John Willard Gibbs even in classical statistical mechanics to resolve the famous Gibbs paradox. Even though this principle has been with us for a very long time, unfortunately, its full significance was not appreciated till 1977 and that is how one missed the possibility of fractional statistics in two dimensions for all these years.

Following Leinaas and Myrheim [8], let us enquire about the configuration space of a system of identical particles? Normally one considers the full phase space in statistical mechanics but it turns out that configuration space is enough for this discussion. Suppose one particle space is  $X$ . Then what is the configuration space of  $N$  identical particles? The Naive answer is  $X^N$ , which, even though true locally, is *not correct* globally. Why? The reason is, since the particles are strictly identical, hence there is no distinction between the points in  $X^N$  that differ only in the ordering of the particle coordinates. For example, consider the point

$$\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) \quad (2)$$

in  $X^N$  where  $\mathbf{x}_i \in X$  for  $i = 1, 2, \dots, N$ . Now consider another point  $\mathbf{x}'$  in  $X^N$  which is obtained from  $\mathbf{x}$  by the permutation  $p$  of the particle indices i.e.

$$\mathbf{x}' = P(\mathbf{x}) = (\mathbf{x}_{p^{-1}(1)}, \dots, \mathbf{x}_{p^{-1}(N)}). \quad (3)$$

Clearly, both describe the *same* physical configuration of the system. Thus the true configuration of the  $N$ -particle system is *not*  $X^N$  but it is the space  $X^N/S_N$  which is obtained by identifying points in  $X^N$  that represent the same physical configuration, i.e. it is obtained from  $X^N$  by dividing out by the action of the symmetry group  $S_N$ . Note that  $S_N$  is a discrete, finite group obtained by permutation of  $N$  identical particles. As a result, the space  $X^N/S_N$  is locally isomorphic to  $X^N$  except at its singular points. However, the global properties of the two spaces are very different. Whereas  $X^N$  has only regular points when  $X$  is regular, those points in  $X^N/S_N$  which correspond to a coincidence of the positions of two or more particles are in fact singular points of  $X^N/S_N$ . Thus to calculate the configuration space of identical particles, such singular points must be excluded by say hard-core constraint so that we can determine if two particles have been exchanged or not. This of course does not make much difference classically. However, in the quantum case the global properties of the configuration space are of deep significance and this results in the possibility of fractional statistics. It is worth emphasizing that this is the crux of the whole matter and it is this fact which was missed for about fifty years!

It turns out that the removal of such singular point in two space dimensions makes the space multiply connected while for three and higher space dimensions it is still doubly connected. That is why, in two dimensions it is possible to define paths that wind around the origin an arbitrary number of times counted with orientation. As a consequence, when one quantizes a system of identical particles then one can show that in two dimensions it is possible to consistently assign *any value* to the phase arising due to the exchange of two identical particles. Since in two dimensions one can distinguish the clockwise winding from the anti-clockwise winding, hence without any loss of generality one can assign the phases  $e^{+i\theta}$  and  $e^{-i\theta}$  respectively, in the case of the anti-clockwise and the clockwise windings.

At this point, it may be worthwhile to mention few key properties of anyons.

1. Anyons must necessarily violate the discrete symmetries of parity (P) and time reversal (T) if  $0 < \theta < \pi$  since the clockwise and the anti-clockwise windings have different phase factors.
2. Anyons are sort of in between the bosons and the fermions i.e. the repulsion between two anyons in the ground state monotonically increases as  $\theta$  goes from 0 to  $\pi$  with there being no repulsion between two bosons. Thus, in a sense, anyons are closer to the fermions than to the bosons since all of them will satisfy a generalized form of Pauli exclusion principle.

3. It turns out that whereas the permutation group which is at the heart of the Bose-Einstein and the Fermi-Dirac statistics, it is the braid group which is at the heart of the fractional statistics. In particular, whereas there are two one dimensional representations of the permutation group (the identical one and the alternating one, corresponding to the Bose-Einstein and Fermi-Dirac statistics respectively), the braid group admits a continuous parameter family of one dimensional representations which one usually identifies with the parameter  $\theta$  which characterizes fractional statistics.
4. Is there a relation between the anyonic statistics and the parastatistics ? The answer is *no*. They are built on two different structures i.e. whereas the Parastatistics corresponds to the higher dimensional representation of the permutation group while anyons correspond to the one dimensional representation of the braid group.

## 2.1 Quantum Statistics in One Dimension

Since we have been talking about the possible quantum statistics in various dimensions, hence it may be worthwhile to also talk about the various possibilities in one dimension. Recall that the notion of the spin does not exist in one dimension since there is no axis to rotate about in that case. Similarly the concept of the quantum statistics is not uniquely defined in one dimension since the position of two particles cannot be interchanged without their passing through one another. As a result, the intrinsic statistics is inextricably mixed up with the local interactions. In fact this ambiguity is at the heart of the bosonization technique which allows the same particle to be represented alternatively by a boson or a fermion field. If, however, statistics is defined in terms of the exclusion principle rather than the exchange of identical particles, then it is possible to define quantum statistics in even one dimension [10].

## 3 Introduction to Chern-Simons Term

We now want to understand how anyons occur in field theory. It turns out that this is possible provided the CS term or its incarnation, the Hopf term are present. It may therefore be worthwhile to first introduce the CS term (in 2+1 dimensions) and discuss its various properties [11].

### 3.1 What is Chern-Simons Term?

Consider the Lagrangian density for classical electrodynamics in 3+1 dimensions as given by

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \bar{\psi}(i\gamma_\mu D^\mu - m)\psi \quad (4)$$

where  $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$  and  $D_\mu = \partial_\mu - ieA_\mu$  is the covariant derivative. This Lagrangian is invariant under the local gauge transformation

$$\psi(x) \rightarrow e^{ie\alpha(x)}\psi(x), \quad A_\mu(x) \rightarrow A_\mu(x) + \partial_\mu\alpha(x). \quad (5)$$

Similarly, for massless fermions ( $m=0$ ), this Lagrangian is also invariant under the (global) chiral transformation

$$\psi(x) \rightarrow e^{i\gamma_5\beta}\psi(x), \quad A_\mu(x) \rightarrow A_\mu(x). \quad (6)$$

The naive expectation was that, these two symmetries i.e. the gauge and the chiral symmetries, which are valid at the classical level, will continue to hold good even in the quantum theory. As a consequence, one expected that the vector and the axial vector currents  $j_\mu = \bar{\psi}\gamma_\mu\psi$  and  $j_\mu^5 = \bar{\psi}\gamma_\mu\gamma_5\psi$  which are conserved at the classical level, will continue to remain conserved even in the quantum theory. It has however, been shown that this is not so. There is *no* regularization which can simultaneously preserve both these symmetries at the quantum level. Because of the unexpected result, it was called an anomaly at that time (and unfortunately even today it is called so), even though the correct name should have been quantum mechanical symmetry breaking.

Remarkably, the entire effect comes only from one loop diagram and two and higher loops do not contribute to the anomaly. In view of our strong faith in the gauge symmetry, one therefore says that it is the chiral symmetry which is broken by the one loop quantum corrections. In particular, there is a gauge singlet (axial) anomaly in any even dimension,  $(2n)$  so that the divergence of the gauge singlet axial current, even for massless fermions, is non-zero and proportional to the corresponding Chern-Pontryagin (CP) density  $P_{2n}$  in that (even) dimension  $2n$  i.e.

$$\partial^\mu j_\mu^5(x) \propto P_{2n}. \quad (7)$$

It is also well known that the CP Density can always be written as a total divergence

$$P_{2n} = \partial_\mu \Lambda^\mu, \quad \mu = 0, 1, 2, \dots, 2n-1. \quad (8)$$

The object  $\Lambda^\mu$ , for a particular value of  $\mu$  (say  $\mu = 2n-1$ ) naturally lives in odd  $(2n-1)$  dimensions and is known as the CS density in that dimension. Thus, whereas the CP density lives in even space-time dimensions, the CS density lives in odd space-time dimensions. For example, the gauge singlet anomaly in 3+1 dimensional quantum electrodynamics is given by

$$\partial^\mu j_\mu^5 = \frac{e^2}{2\pi} \varepsilon_{\mu\nu\lambda\sigma} F^{\mu\nu} F^{\lambda\sigma} = \frac{e^2}{\pi} \partial^\mu (\varepsilon_{\mu\nu\lambda\sigma} A^\nu F^{\lambda\sigma}) \quad (9)$$

so that the Abelian CS term in 2+1 dimensions is given by

$$J_{CS} = \int \mathcal{L}_{CS} d^3x \propto \int d^3x \varepsilon_{\nu\lambda\sigma} A^\nu F^{\lambda\sigma}. \quad (10)$$

Throughout this book we shall mainly be concerned with this CS term or its non-Abelian generalization. Let us therefore discuss in some detail the various properties of this term.

### 3.2 Gauge Invariant Mass Term

Let us consider pure electrodynamics in the presence of the Chern-Simons term in 2+1 dimensions [12, 13]

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + \frac{\mu}{4} \varepsilon^{\mu\nu\lambda} F_{\mu\nu} A_\lambda. \quad (11)$$

Since the mass dimension of  $A_\mu$  is  $1/2$ , hence it follows that the parameter  $\mu$  has the dimension of mass. The field equation following from this Lagrangian can be written as

$$(g^{\mu\nu} + \frac{1}{\mu} \varepsilon^{\mu\nu\alpha} \partial_\alpha) {}^*F_\nu = 0 \quad (12)$$

where  ${}^*F_\nu$  is the dual field strength which is a vector in 2+1 dimensions i.e.

$${}^*F_\nu = \frac{1}{2} \varepsilon_{\nu\alpha\beta} F^{\alpha\beta}; \quad F_{\mu\nu} = \varepsilon_{\mu\nu\alpha} {}^*F^\alpha. \quad (13)$$

We thus find that, unlike the CP term which has only a nontrivial topology but no dynamics (being a total divergence), the CS term has nontrivial topology as well as dynamics in it. On operating by  $(g_{\beta\eta} - \frac{1}{\mu} \varepsilon_{\beta\eta\delta} \partial^\delta)$  to Eq. (12), we get

$$(\partial_t^2 - \nabla^2 + \mu^2) {}^*F_\beta = 0 \quad (14)$$

which clearly shows that the gauge field excitations are massive with the gauge field mass  $\mu$  being the coefficient of the CS term. We have thus shown that the CS term when added to the Maxwell term, acts as the *gauge invariant gauge field mass term*. It is worth adding that this remarkable property of having a gauge invariant mass term for the gauge field in the action itself is very special to 2+1 dimensions.

### 3.3 Behavior Under C, P, and T

Let us consider the behaviour of the CS term as well as the Dirac Lagrangian

$$\mathcal{L}_D = i\bar{\psi}(\gamma_\mu \partial^\mu - m)\psi \quad (15)$$

under the discrete transformations  $C$  (charge conjugation),  $P$  (parity) and  $T$  (time reversal). Here,  $\psi$  is a two component spinor with mass  $m(0)$  and the mass dimension of  $\psi$  is 1. We use the following two-dimensional realization of the Dirac algebra

$$\gamma^0 = \sigma^3; \quad \gamma^1 = i\sigma^1, \quad \gamma^2 = i\sigma^2, \quad (16)$$

$$\gamma^\mu \gamma^\nu = g^{\mu\nu} - i\varepsilon^{\mu\nu\alpha} \gamma_\alpha; \quad g^{\mu\nu} = \text{diag.}(1, -1, -1) \quad (17)$$

where  $\sigma^i$  are the usual Pauli matrices.

It is easily shown that under charge conjugation

$$CA_\mu C^{-1} = -A_\mu, \quad C\psi C^{-1} = \sigma^1 \psi^\dagger \quad (18)$$

so that the action is invariant under  $C$ . On the other hand, under parity transformation, the gauge and the Fermi fields transform as follows

$$PA^{0,2}(t, \mathbf{r})P^{-1} = A^{0,2}(t, \mathbf{r}'), \quad PA^1(t, \mathbf{r})P^{-1} = -A^1(t, \mathbf{r}'), \quad (19)$$

$$P\psi(t, \mathbf{r})P^{-1} = \sigma^1 \psi(t, \mathbf{r}'). \quad (20)$$

Note that in 2+1 dimensions, the parity transformation is somewhat unusual i.e.  $\mathbf{r} = (x, y)$ ,  $\mathbf{r}' = (-x, y)$  (or  $(x, -y)$ ). On the other hand,  $(-x, -y)$  corresponds to rotation (and not space reflection). As a result, we find that the mass terms for both the Fermi and the gauge fields (i.e.  $m\bar{\psi}\psi$  and the CS term) are not invariant under parity. Similarly, time-inversion changes the signs of both the mass terms since

$$TA^0(t, \mathbf{r})T^{-1} = A^0(-t, \mathbf{r}), \quad TA(t, \mathbf{r})T^{-1} = -A(-t, \mathbf{r}), \quad (21)$$

$$T\psi(t, \mathbf{r})T^{-1} = \sigma^2 \psi(-t, \mathbf{r}). \quad (22)$$

Thus, both the CS term as well as the fermion mass term,  $m\bar{\psi}\psi$  are non-invariant under  $P$  as well as  $T$ . However, they are invariant under the combined operation  $PT$  and hence the  $CPT$  symmetry is still valid. Note that in 3+1 dimensions though,  $m\bar{\psi}\psi$  is invariant under  $P, C$  and  $T$  separately.

Finally, let us talk about the photon spin. One can show that the CS photon spin is 1(-1) if CS mass  $\mu_0(< 0)$  while the spin of the massless photon is zero. Further, in either case, the photon has only one degree of freedom.

### 3.4 Coleman - Hill Theorem

It turns out that because of the  $P$  and  $T$  violating but gauge invariant CS term, the most general form for the vacuum polarization tensor consistent with Lorentz and gauge invariance is more general than in other dimensions i.e.

$$\Pi_{\mu\nu}(k) = (k^2 g_{\mu\nu} - k_\mu k_\nu) \Pi_1(k^2) - i\varepsilon_{\mu\nu\lambda} k^\lambda \Pi_2(k^2). \quad (23)$$

Note that the second term on the right hand side is odd under  $P$  and  $T$ . It is clear that any  $P$  and  $T$  violating interaction will contribute to  $\Pi_2(k^2)$ . For example, the fermion mass term which violates both  $P$  and  $T$ , does contribute to  $\Pi_2(k^2)$  at one loop. Remarkably enough, it was discovered that at two loops, however, there is no contribution to  $\Pi_2(0)$  and hence to Chern-Simons mass [11]. Inspired by this result, Coleman and Hill [14] have in fact proved under very general conditions that  $\Pi_2(0)$  receives no contribution from two and higher loops in any gauge and Lorentz invariant theory including particles of spin 1 or less (An open question is whether this is also valid for higher

spin theories, specially spin-3/2). They only require that the matter fields be massive so that one does not have to worry about the infrared problems. Further, they also assume that no part of the free electro-magnetic Lagrangian density is hiding in the matter part of the Lagrangian. It may be noted that their result is valid even for non-renormalizable interactions in the presence of the gauge and Lorentz invariant regularization.

Coleman and Hill also claimed that at one loop, the only contribution to  $\Pi_2(0)$  can come from the fermion loop. This is, however, incorrect. In particular, there is no reason why  $P$  and  $T$  violating interactions involving spin-0 or spin-1 particles should not contribute to  $\Pi_2(0)$  at one loop. In fact, it has been shown that the parity violating spin-0 [15] as well as spin-1 interactions [16] do contribute to  $\Pi_2(0)$  at one loop.

### 3.5 Magneto-Electric Effect

There are many crystals in nature like chromium oxide, which show the magneto-electric effect i.e., they also get magnetically polarized in an electric field and electrically polarized in a magnetic field [17, 18]. It is well known that this effect depends upon having a  $CP$ -asymmetric medium. Mathematically, the signal for the magneto-electric effect in 2+1 dimensions is that the relation between the excitation fields  $\mathbf{D}$  and  $\mathbf{H}$  and  $\mathbf{E}$  and  $\mathbf{B}$  is modified to

$$D_i = \chi_{ij}^{(e)} E_j + \chi_i^{(em)} B; \quad H = \chi^{(m)} B + \chi_i^{(me)} E_i. \quad (24)$$

It has been shown [19] that the vacuum of the 2 + 1 dimensional quantum electrodynamics with CS term also shows the magneto-electric effect. In particular, it has been shown that both  $\chi_i^{(em)}$  and  $\chi_i^{(me)}$  are non-zero and proportional to  $k_i \Pi_2(k^2)$ . Of course this is not really surprising if one remembers that the CS term violates the discrete symmetries  $P$  and  $T$ .

### 3.6 Chern-Simons Term by Spontaneous Symmetry Breaking

We have seen above that the CS term provides mass to the gauge field. Now, usually the gauge field mass is generated by spontaneous symmetry breaking; hence it is worth enquiring whether the CS term can also be generated by spontaneous symmetry breaking. The answer to the question is yes [20]. This is because, unlike other dimensions, in the 2 + 1 case, one can have a more general definition of the covariant derivative. In particular, it is easily seen that

$$\mathcal{D}_\mu \psi = (\partial_\mu - ieA_\mu - ig\varepsilon_{\mu\nu\lambda} F^{\nu\lambda})\psi \quad (25)$$

also transforms as a covariant derivative, since the field strength  $F^{\nu\lambda}$  by itself is gauge invariant. Obviously, the same thing is also true for a spin-0 charged scalar field. Now consider the following generalized Abelian Higgs model in 2 + 1 dimensions

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{1}{2}(\mathcal{D}_\mu \phi)^*(\mathcal{D}^\mu \phi) - \alpha(|\phi|^2 - a^2)^2 \quad (26)$$

where the generalized covariant derivative is as given by Eq. (25). On expanding the term  $\frac{1}{2}(\mathcal{D}_\mu \phi)^*(\mathcal{D}^\mu \phi)$ , we have

$$\begin{aligned} \frac{1}{2}(\mathcal{D}_\mu \phi)^*(\mathcal{D}^\mu \phi) &= \frac{1}{2}(\partial_\mu + ieA_\mu)\phi^*(\partial^\mu - ieA^\mu)\phi + \frac{g^2}{4}F_{\mu\nu}F^{\mu\nu}|\phi|^2 \\ &\quad + ig^*F_\mu(\phi^*\partial^\mu\phi - \phi\partial^\mu\phi^*) + eg\varepsilon_{\mu\nu\lambda}(\partial^\mu A^\nu)A^\lambda|\phi|^2 \end{aligned} \quad (27)$$

so that if  $\phi$  acquires a nonzero vacuum expectation value then the Abelian CS term is generated from the last term of this equation. Clearly a similar mechanism should also work for the non-Abelian case, but technically it is a tougher problem since one also has to generate the non-linear term.



### 3.7 Lorentz Invariance From Gauge Invariance

One of the remarkable properties of the Abelian CS term is that in this case the Lorentz invariance of the action automatically follows from the gauge invariance. In contrast, notice that the most general form of the gauge invariant Maxwell Lagrangian in classical electrodynamics in  $3 + 1$  dimensions is

$$\mathcal{L} = \mathbf{E}^2 + a\mathbf{B}^2. \quad (28)$$

It is only the demand of the Lorentz invariance which tell us that  $a = -1$  (In the  $2 + 1$  case,  $B$  is a pseudo scalar but the same argument is still valid). On the other hand, if one writes the CS action as

$$I_{CS} = \int d^3x [\epsilon_{ij} E^i A^j + aBA^0], \quad (29)$$

then the demand of the invariance of  $I_{CS}$  under the gauge transformation  $A_\mu \rightarrow A_\mu + \partial_\mu \alpha$  fixes  $a$  and uniquely gives us the CS action which is automatically also Lorentz invariant.

### 3.8 Quantization of Chern-Simons Mass

Let us now discuss the CS term in the non-Abelian gauge theories. We shall mention only those properties which are special to the non-Abelian CS term. To begin with, notice that the non-Abelian CS term has an extra term compared to the Abelian case i.e.

$$I_{na}^{(CS)} = \frac{\mu}{4} \int d^3x \epsilon^{\mu\nu\lambda} \text{tr}(F_{\mu\nu} A_\lambda - \frac{2}{3} A_\mu A_\nu A_\lambda) \quad (30)$$

where  $A_\mu$  and  $F_{\mu\nu}$  are matrices

$$A_\mu = gT^a A_\mu^a; F_{\mu\nu} = gT^a F_{\mu\nu}^a = \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu]. \quad (31)$$

Here,  $T^a$  are the representation matrices of the gauge group  $G$  satisfying

$$[T^a, T^b] = f^{abc} T^c \quad (32)$$

where  $f^{abc}$  are the structure constants of the group. In the case of  $SU(2)$ ,  $T^a = \tau^a/2i$ .

Let us now consider a non-Abelian gauge theory with the Chern-Simons term as given by

$$\mathcal{L}_{na} = \frac{1}{2g^2} \text{tr}(F^{\mu\nu} F_{\mu\nu}) - \frac{\mu}{2g^2} \epsilon^{\mu\nu\lambda} \text{tr}(F_{\mu\nu} A_\lambda - \frac{2}{3} A_\mu A_\nu A_\lambda) \quad (33)$$

As in the Abelian case, it is easily shown that the CS term provides a gauge invariant gauge field mass  $\mu$ .

As in the Abelian case, the non-Abelian CS Lagrangian density changes by a total derivative under an infinitesimal local gauge transformation so that the corresponding action is invariant under such a gauge transformation. However, the CS action is not invariant under finite (also called homotopically non-trivial, or those which are not continuously deformable to the identity) gauge transformations as given by

$$A_\mu \rightarrow U^{-1} A_\mu U + U^{-1} \partial_\mu U. \quad (34)$$

As a result, one finds that the action corresponding to the Lagrangian (33) transforms as follows

$$\begin{aligned} I_{na} &\rightarrow I_{na} + \mu \int d^3x \epsilon^{\mu\nu\lambda} \text{tr} \left( \partial_\nu [A_\mu (\partial_\lambda U) U^{-1}] \right) \\ &+ \frac{\mu}{3} \int d^3x \epsilon^{\mu\nu\lambda} \text{tr} \left[ (\partial_\mu U) U^{-1} (\partial_\nu U) U^{-1} (\partial_\lambda U) U^{-1} \right]. \end{aligned} \quad (35)$$

Let us consider those gauge transformations which tend to the identity at temporal and spatial infinity so as to avoid a convergence problem i.e.

$$U(X) \xrightarrow{x \rightarrow \infty} I. \quad (36)$$

It is now easily seen that the gauge field dependent surface integral in Eq. (35) vanishes. However, the last term in the integral is non-zero. It can be converted to a surface integral once the integrand is rewritten as a total derivative. This can be made manifest by using an explicit parameterization for  $U$ . For example, in the case of  $SU(2)$  (more generally, we choose  $SU(2)$  sub-group of the gauge group  $G$ ; for reasons that will be clear soon), one can make use of the exponential parameterization  $U(X) = \exp(i\sigma^a\theta^a(x))$ . In this way one can show that under large gauge transformations,  $I_{na}$  is not invariant but transforms as

$$I_{na} \rightarrow I_{na} + \frac{8\pi^2\mu}{g^2}\omega(U) \quad (37)$$

where

$$\omega(U) = \frac{1}{24\pi^2} \int d^3x \varepsilon^{\mu\nu\lambda} \text{tr} \left[ (\partial_\mu U)U^{-1}(\partial_\nu U)U^{-1}(\partial_\lambda U)U^{-1} \right] \quad (38)$$

is the winding number of the gauge transformation  $U$ . In particular, if the gauge group  $G$  is such that the third homotopy group of  $G$  is non-trivial i.e.

$$\pi_3(G) = Z \quad (39)$$

where  $Z$  is the additive group of integers, then under these so called large gauge transformations, the action transforms as

$$I_{na} \rightarrow I_{na} + \frac{8\pi^2\mu}{g^2}m \quad (40)$$

where  $m$  is an integer. Note in particular, that Eq. (39) is true for any gauge group  $G$  of which  $SU(2)$  is a sub-group. However, in the path integral formulation, the action itself may or may not be gauge invariant but, it is the exponential of the action ( $\exp(iI_{na})$ ) which should be gauge invariant. In this way we conclude that the non-Abelian gauge theory with the CS term does not make sense in  $2+1$  dimensions unless the CS mass  $\mu$  is quantized [13] in units of  $g^2/4\pi$  i.e. ( $n = 0, \pm 1, \pm 2, \dots$ )

$$\frac{8\pi^2\mu}{g^2} = 2\pi n \quad \text{or} \quad \mu = \frac{g^2}{4\pi}n. \quad (41)$$

This mass quantization is reminiscent of the famous Dirac quantization in the case of magnetic monopole. An important question to address is whether the quantization condition (41) is respected by the quantum corrections. This issue was considered by Pisarski and Rao [21] for the case of a pure gauge theory (i.e. without any matter field). They found that the quantization is indeed preserved to one loop; however, the integer on the right hand side of Eq. (41) is shifted by  $N$  in case the gauge group  $G = SU(N)$ . Subsequently, it has been shown that there are no further corrections from two and higher loops in the limit of the pure CS gauge theory [22].

How does the quantization condition modify in the presence of the matter fields? It has been shown that so long as the scalar field does not break the non-Abelian gauge symmetry, then the quantization condition remains unaltered. The massive fermions, of course, modify the quantization condition [21]; the right hand side of Eq. (41) being shifted by  $\frac{m_f}{|m_f|}T_R$ , where  $T_R$  is the Casimir generator for the gauge group  $G$  (i.e.  $\text{tr}(T^a T^b) = -\delta^{ab}T_R$ ), in case the fermions are in the fundamental representation of the gauge group  $G$ . Thus the quantization is preserved so long as  $T_R$  is an integer.

Much more interesting is the case of partial (spontaneous symmetry) breaking of a non-Abelian gauge symmetry. In this case it has been shown that if the non-Abelian gauge symmetry  $SU(N)$  is spontaneously broken to say  $SU(M) \otimes U(1)$  (or even several  $U(1)$ 's), then the one-loop radiative correction to the right hand side of the quantization condition (41) [23] arises purely from the unbroken non-Abelian sector in question, the orthogonal  $U(1)$  sector makes no contribution. This implies that the coefficient of the CS term is a discontinuous function over the phase diagram of the theory.

### 3.9 Parity Anomaly

Is our entire discussion about the CS term merely of academic interest ? Put differently, some one might argue that since the CS term violates both the parity and the time reversal invariance symmetries, why should one, in the first place, add such a term to the action ? The answer to this question, at least in the non-Abelian gauge theories, is that even if one does not add the CS term to the action at the tree level, it is automatically generated by the one loop radiative corrections due to the so called parity anomaly [24]. In particular, consider the action

$$I[A_\mu, \psi] = \int d^3x \left[ \frac{1}{2g^2} \text{tr}(F_{\mu\nu} F^{\mu\nu}) + i\bar{\psi}\gamma_\mu(\partial^\mu - ieA^\mu)\psi \right] \quad (42)$$

for an odd number of massless doublet of fermions in the fundamental representation coupled to  $SU(2)$  gauge fields (more generally any gauge group  $G$  of which  $SU(2)$  is a sub-group so that Eq. (39) is satisfied; and the fermions are required only to be in the fundamental representation).

This action is invariant under the gauge transformations (both large and small) as well as the discrete transformations of parity (P) and time reversal invariance (T). However, the effective action  $I_{eff}[A]$ , obtained by integrating out the fermionic degrees of freedom, violates one of the two symmetries. In other words, there is *no* regularization which can simultaneously maintain the invariance of  $I_{eff}[A]$  under the large gauge transformations as well as  $P$  and  $T$ . In view of the tremendous success of the gauge principle, one usually maintains the gauge invariance at the cost of the parity and the time reversal invariance by simply adding the CS term to the action (alternately one can also regulate it by using the  $P$  and  $T$  violating Pauli-Villars regularization). In this way, one finds that the CS term is induced by the radiative corrections even if it is absent at the tree level. This is very similar to the way the CP term is induced in even dimensions due to the gauge singlet (chiral) anomaly.

### 3.10 Topological Field Theory

One of the most remarkable property of the CS action is that it depends only on the antisymmetric tensor  $\varepsilon_{\mu\nu\lambda}$  and not on the metric tensor  $g_{\mu\nu}$ . As a result, the CS action in the flat and the curved space is the same. Hence, the CS action, in both the Abelian and the non-Abelian cases, is an example of the topological field theory [25]. It might be mentioned here, that the topological field theories give a natural framework for understanding the Jones polynomials of the Knot theory in terms of three dimensional terms. Further, these theories have shed new light on conformal field theories in two space-time dimensions.

Finally, the gravitational Chern-Simons term has also been considered [13] and shown to have some remarkable properties. In particular, whereas the massless Einstein theory in 2+1 dimensions is trivial, it acquires a propagating, massive, spin-2 degree of freedom when the CS term is present. Further, even though this topological term has third time derivative dependence, yet the theory is ghost-free and unitary and one has a consistent quantum theory. The contribution of the topological mass term to the field equations also has a natural geometric significance: it is the three dimensional analogue of the Weyl tensor.

## 4 Charged Vortex as Anyon in Field Theories

In the last section, we have discussed in detail the various properties of the CS term. In this section, we demonstrate the most dramatic effect of this term i.e. the existence of charged vortex solutions thereby providing us with a relativistic model for the charged (extended) anyons.

Before we discuss the charged vortex solutions, it might be worthwhile to mention how such solutions were historically discovered. A long time ago, Abrikosov [26] wrote down the electrically neutral vortex solutions in the Ginzburg-Landau theory which is a mean-field theory of superconductivity. Subsequently, these vortices were experimentally observed in the type-II superconductors. Nielsen and Olesen [27] rediscovered these solutions in the context of the Abelian

Higgs model which is essentially a relativistic generalization of the Ginzburg-Landau theory. These people were looking for string-like objects in relativistic field theory. It turns out that these vortices have finite energy per unit length in  $3 + 1$  dimensions (i.e. finite energy in  $2 + 1$  dimensions as the vortex dynamics is essentially confined to the  $x$ - $y$  plane), quantized flux, but are electrically neutral and have zero angular momentum. Subsequently, Julia and Zee [28] showed that the  $SO(3)$  Georgi-Glashow model which admits t'Hooft-Polyakov monopole solution, also admits its charged generalization i.e. the dyon solution with finite energy and finite, non-zero, electric charge. It was then natural for them to enquire whether the Abelian Higgs model, which admits neutral vortex solutions with finite energy (in  $2+1$  dimensions), also admits its charged generalization or not. In the appendix of the same paper, Julia and Zee discussed this question and showed that the answer is *no* i.e. unlike the monopole case, the Abelian Higgs model does not admit charged vortices with finite energy and finite and non-zero electric charge. More than ten years later, Samir Paul and I [29] showed that the Julia-Zee negative result can be overcome if one adds the CS term to the Abelian Higgs model. In particular, we showed that the Abelian Higgs model with CS term in  $2+1$  dimensions admits charged vortex solutions of finite energy and quantized, finite, Noether charge as well as flux. As an extra bonus, it was found that these vortices also have non-zero, finite angular momentum which is in general fractional. This strongly suggested that these charged vortices could in fact be charged anyons which was subsequently rigorously shown by Fröhlich and Marchetti [30].

Strictly speaking, what one has obtained are the charged soliton solutions and not the vortex solutions, but because of the close connection with the neutral vortex solutions, one has continued to call them as charged vortices rather than charged solitons.

Consider an Abelian Higgs model with CS term as given by

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{1}{2}(D_\mu\phi)^*(D^\mu\phi) - C_4(|\phi|^2 - \frac{C_2}{2C_4})^2 + \frac{\mu}{4}\varepsilon_{\mu\nu\lambda}F^{\mu\nu}A^\lambda \quad (43)$$

where  $\mu$  is the Chern-Simons mass,  $\phi$  denotes complex scalar field and  $D_\mu\phi$  is the covariant derivative i.e.

$$D_\mu\phi = (\partial_\mu - ieA_\mu)\phi. \quad (44)$$

Here  $\phi, A_\mu$  as well as the gauge coupling constant  $e$  have mass dimension of  $1/2$  while  $C_4$  and  $C_2$  have mass dimensions of 1 and 2 respectively. In order to obtain the charged vortex solutions, let us consider the following *ansatz*

$$\mathbf{A}(\mathbf{x}, t) = -e_\theta C_0 \frac{(g(r) - n)}{r}, \quad \phi(\mathbf{x}, t) = C_0 f(r) e^{in\theta}, \quad A_0(\mathbf{x}, t) = C_0 h(r) \quad (45)$$

where  $g(r), h(r), f(r)$  are the dimension-less fields,  $r$  is the dimension-less length, while  $C_0$  has mass dimension of  $1/2$  i.e.

$$r = eC_0\rho, \quad C_0 = \sqrt{C_2/2C_4}. \quad (46)$$

Note that  $\rho$  and  $\theta$  are related to  $x$  and  $y$  by  $\rho = \sqrt{x^2 + y^2}$  and  $\tan\theta = y/x$ . It turns out that even though the Lagrangian (43) has so many parameters, the dynamics essentially depends on two dimension-less variables,  $\delta$  and  $\lambda$  defined by

$$\lambda = \sqrt{8C_4/e^2}, \quad \delta = \mu/eC_0. \quad (47)$$

The field equations which follow from here are

$$g''(r) - \frac{1}{r}g'(r) - gf^2 = \delta rh'(r) \quad (48)$$

$$h''(r) + \frac{1}{r}h'(r) - hf^2 = \frac{\delta}{r}g'(r) \quad (49)$$

$$f''(r) + \frac{1}{r}f'(r) - \frac{g^2f}{r^2} + \frac{\lambda^2}{2}f(1 - f^2) = -fh^2 \quad (50)$$

where  $g'(r) \equiv dg(r)/dr$ . The corresponding field energy can be shown to be

$$E_n = \pi C_0^2 \int_0^\infty r dr \left[ \frac{1}{r^2} \left( \frac{dg}{dr} \right)^2 + \left( \frac{df}{dr} \right)^2 + \left( \frac{dh}{dr} \right)^2 + h^2 f^2 + \frac{g^2 f^2}{r^2} + \frac{\lambda^2}{4} (1 - f^2)^2 \right] \quad (51)$$

Several remarks are in order at this stage.

1. As expected, in the limit  $h = 0$  (i.e.  $A_0 = 0$ ) and  $\delta = 0$  (i.e.  $\mu = 0$ ) the field equations reduce to those of the neutral vortex case. From the Gauss law Eq. (49) it also follows that if  $\delta$  (i.e.  $\mu$ ) is non-zero, then  $A_0$  must also be non-zero thereby justifying the *ansatz* (45).
2. The boundary conditions for finite energy solutions are

$$\lim_{r \rightarrow \infty} f(r) = 1, h(r) = 0 = g(r) \quad (52)$$

$$\lim_{r \rightarrow 0} f(r) = 0, g(r) = n, h(r) = \beta \quad (53)$$

where  $\beta$  is an arbitrary number while  $n = 0, \pm 1, \pm 2, \dots$ .

3. From these boundary conditions it immediately follows that the magnetic flux is quantized in units of  $2\pi/e$  i.e.

$$\Phi \equiv \int B d^2x = -\frac{2\pi}{e} \int_0^\infty r dr \left( \frac{1}{r} \frac{dg}{dr} \right) = \frac{2\pi n}{e}. \quad (54)$$

It may be noted that even for the neutral vortices, the flux is quantized in units of  $\frac{2\pi}{e}$ . The underlying reason for the flux quantization is same in both the cases i.e. both are topological objects with the underlying boundary conditions being such that there is a non-trivial mapping from the space time to the group manifold i.e.  $\pi_1(U(1)) = Z$ , with  $Z$  being the set of integers, forming a group under addition.

4. From the Gauss law Eq. (49), it then follows that these vortices also have a non-zero and finite Noether charge which is quantized in units of  $2\pi\mu/e$ . This is easily seen by noting that in terms of the electric and the magnetic fields, the Gauss law equation can be written as

$$\nabla \cdot \mathbf{E} + \mu B = \rho \quad (55)$$

where  $\rho$  is the Noether charge density. On integrating both sides of this equation, it then follows that

$$Q \equiv \int \rho d^2x = \mu \int B d^2x = \frac{2\pi\mu}{e} n. \quad (56)$$

Note that  $\int \nabla \cdot \mathbf{E} d^2x = 0$ , since, because of the Higgs mechanism, both  $\mathbf{E}$  and  $B$  fall off exponentially at long distances. This is probably for the first time that the quantization of the Noether charge has followed from purely topological considerations. In a sense, relation (56) can be looked upon as the (2+1)-analogue of the Witten effect [31]. Let us recall the work of Witten who had shown that in the presence of the  $CP$  and  $T$  violating  $CP$  term, the t'Hooft-Polyakov monopole acquires electric charge whose fractional part is proportional to the coefficient of the  $CP$  term. It must however be remembered that whereas the Witten effect is purely a quantum mechanical effect, in our case, the vortices acquire a non-zero charge at the classical level itself due to the presence of the CS term.

5. It is also clear from here that in the Abelian Higgs model (without the CS term), one cannot have vortices having simultaneously the finite energy as well as the finite, non-zero Noether charge. The point is, in the absence of the CS term, the Gauss law Eq. (55) gives on integration

$$Q \equiv \int \rho d^2x = \int \nabla \cdot \mathbf{E} d^2x. \quad (57)$$

The only way  $Q$  can be non-zero and finite is if there is a non-zero contribution to the integral around  $r \rightarrow 0$  i.e. if  $\mathbf{E} \rightarrow 1/r$  as  $r \rightarrow 0$ . But in that case, the electrical field energy  $\int \mathbf{E}^2 d^2x$  diverges logarithmically [28].

6. The energy-momentum tensor  $T_{\mu\nu}$  for this model can be obtained by varying the curved space form of the action with respect to the metric

$$T_{\mu\nu} = \frac{1}{2}(D_\mu\phi)^*(D_\nu\phi) + \frac{1}{2}(D_\nu\phi)^*D_\mu\phi - g_{\mu\nu}(\mathcal{L} - \frac{\mu}{4}\epsilon_{\alpha\beta\gamma}F^{\alpha\beta}A^\gamma) \quad (58)$$

where the Lagrangian  $\mathcal{L}$  is as given by Eq. (43). Note that the CS term, being independent of the metric tensor  $g_{\mu\nu}$ , does not contribute to the energy momentum tensor  $T_{\mu\nu}$ . Using this  $T_{\mu\nu}$  and the field equations, the angular momentum carried by the charged vortices can be shown to be

$$J \equiv \int d^2x \epsilon^{ij} x_i T_{0j} = -\frac{nQ}{2e} = -\frac{\pi\mu}{e^2}n^2 = -\frac{Q\Phi}{4\pi}. \quad (59)$$

Thus, unlike the neutral vortices, the angular momentum of the charged vortices is non-zero and is solely determined by their charge and flux. Besides, the angular momentum of  $n$  superimposed charged vortices is  $n^2$  and not  $n$  times the angular momentum of a single vortex. Further, since the CS mass  $\mu$  is not quantized in the Abelian case, hence this angular momentum  $J$  can in general take any fractional value. This strongly suggests that these charged vortices are charged anyons. Fröhlich and Marchetti [30] have in fact rigorously proved that these charged vortices are charged anyons. In particular, they constructed quantum one vortex operator and then evaluated the phase acquired when one such vortex is slowly taken round the other. They also show that the charged vortices cannot be localized in bounded regions but are localized in space-like cones in three-dimensional Minkowski space-time [32]. Unfortunately their treatment is rather involved and is beyond the scope of this pedagogical article. Thus the solitons of the Abelian Higgs model with the CS term provides us with a relativistic field theory model for the extended charged anyons.

7. The magnetic moment of these vortices can be computed by using the field equations and one can show that, whereas for the neutral vortices it is equal to the flux  $\Phi (= 2\pi n/e)$ , the charged ones acquire an extra contribution

$$K_z \equiv \int (\mathbf{r} \times \mathbf{J})_z d^2x = \frac{2\pi n}{e} + \frac{2\pi\delta}{e} \int_0^\infty r h(r) dr. \quad (60)$$

#### 4.1 Unusual Higgs Mechanism

One must now solve the field Eqs. (48) to (50) and show the existence of the charged vortex solutions. To date, no analytic solution has been obtained of these field equations. However, it is easily seen that for large  $r$ , the asymptotic values of the gauge and the Higgs fields are reached exponentially fast

$$g(r) = \alpha_\pm \sqrt{r} e^{-\eta_\pm r} + \dots, \quad h(r) = \mp \frac{\alpha_\pm}{\sqrt{r}} e^{-\eta_\pm r} + \dots, \quad (61)$$

$$f(r) = 1 + \beta e^{-\lambda r} + \dots \quad (62)$$

where  $\alpha_\pm$  and  $\beta$  are dimension-less constants while the dimension-less vector meson mass  $\eta_\pm$  is given by

$$\eta_\pm = \sqrt{1 + \frac{\delta^2}{4}} \pm \frac{\delta}{2}. \quad (63)$$

However, it has subsequently been shown that the solution with  $\eta_+$  does not exist for all  $r$ .

On noting that the field Eqs. (48) to (50) are invariant under  $r \rightarrow -r$ , it is easily shown that the behavior of the gauge and the Higgs fields around  $r = 0$  is given by

$$g(r) = n + \alpha_1 r^2 + O(r^4), \quad h(r) = \beta + \alpha_1 \delta \frac{r^2}{2} + O(r^4), \quad (64)$$

$$f(r) = \alpha_2 r^{|n|} + O(r^{|n|+2}). \quad (65)$$

Detailed numerical work has subsequently confirmed the existence of the radially symmetric charged vortex solutions with these boundary conditions [34]. These correspond to  $n$  superimposed vortices. The qualitative behaviour of the charged vortex solution which follows from here is as follows : the magnetic field  $B$  decreases monotonically from its non-zero value at the core of the vortex ( $r = 0$ ) to reach zero as  $r \rightarrow \infty$  with the penetration length  $1/\eta_-$ , while the Higgs field increases from zero at the origin to its vacuum value at infinity with coherence length  $1/\lambda$ . Finally, the electric field  $E_\rho$  which is radial, vanishes both at  $r = 0$  and  $r = \infty$  reaching the maximum in between at some finite  $r$ . It is worth pointing out that as in the quantum Hall effect, for the charged vortex solutions too,  $\mathbf{E}(\equiv E_\rho)$  is at right angles to  $\mathbf{J}(\equiv j_\theta)$  and both in turn are at right angles to  $\mathbf{B}$ .

Why did one obtain two asymptotic solutions for  $g$  and  $h$ , i.e. for the gauge fields  $A_\theta$  and  $A_0$ ? This is because of the unusual nature of the Higgs mechanism in  $2 + 1$  dimensions in the presence of the CS term. Notice that in our case both the Maxwell and the CS terms are present and in addition there is also Higgs mechanism in operation. Clearly such a theory must still propagate only two massive modes. As has been shown in [35], in this case  $\mathcal{L}_{quad}$  corresponds to Proca equation with the CS term. It propagates a self-dual field with two distinct CS type masses and that corresponding to each mass there is one ( $P$  and  $T$  violating) propagating mode. Further, the two masses (in dimension-less form) are precisely  $\eta_\pm$  as given by Eq. (63) thereby explaining the reason for the occurrence of two asymptotic solutions  $\eta_\pm$ .

## 4.2 Vortex-Vortex Interaction

One of the most interesting question is whether these charged vortices can be observed experimentally in some planar system. In this context recall that the neutral (Abrikosov) vortices have been experimentally seen in type-II superconductors. This can be understood from the fact that whereas the vortex-vortex interaction is repulsive in the type-II region ( $\lambda 1$ ), it is attractive in the type-I region of superconductivity. It is thus of great interest to study the charged vortex-vortex interaction and to see when is it repulsive. This has been done both in the perturbation theory (in the CS mass) and by the variational calculation.[34], and in both cases one finds that the charged vortex-vortex interaction is more repulsive than the corresponding neutral case with the extra repulsion coming from the electric field of the charged vortex. For example, when the CS mass is small, then on expanding the charged  $n$ -vortex fields in terms of the corresponding neutral vortex fields it has been shown that

$$E_n(\lambda, \delta) - nE_1(\lambda, \delta) = E_n(\lambda, 0) - nE_1(\lambda, 0) + \frac{(n^2 - n)}{4}\delta^2 + O(\delta^4) \quad (66)$$

so that the charged vortex-vortex interaction is always more repulsive than the corresponding neutral case. For example, for  $\delta = 0.5$ , one finds that the charged vortex-vortex interaction is repulsive even for  $\lambda > 0.45$  (note that in the neutral case the interaction is repulsive only if  $\lambda 1$ ).

## 4.3 Non-Abelian Charged Vortex Solutions

It is clearly of considerable interest to enquire whether the charged vortex solutions obtained above can be embedded in non-Abelian gauge theories with the CS term. The first obvious question is whether such vortices could be topologically stable or not. It is easily seen that if  $G$  is the gauge group of the non-Abelian gauge theory and  $H$  is the sub-group under which the vacuum remains invariant after spontaneous symmetry breaking, then topologically non-trivial vortices are possible only if

$$\pi_1(G/H) \neq 0. \quad (67)$$

In the case of  $SU(N)$  gauge theories, it turns out that no  $Z$ -vortices are possible. However,  $Z_N$ -vortices are possible in case  $H$  is  $Z_N$  since  $\pi_1(SU(N)/Z_N) = Z_N$ . It turns out that at least  $N$  Higgs multiplets are required so that the vacuum is invariant under  $Z_N$  [36]. As a result, only one non-trivial charged vortex is possible in the case of  $SU(2)$  gauge theory with flux  $\Phi = 2\pi/g$ , charge  $Q = \mu\Phi = 2\pi\mu/g$ , and angular momentum  $J = -Q\Phi/4\pi = -\pi\mu/g^2$  where  $g$  is the gauge coupling

constant. But since the CS mass  $\mu$  is quantized in non-Abelian gauge theories having  $SU(2)$  as its sub-group i.e.

$$\mu = \frac{g^2}{4\pi}n, \quad n = 0, \pm 1, \pm 2, \dots \quad (68)$$

and hence the vortex charge is  $gn/2$  i.e. it is quantized in units of  $g/2$  while the angular momentum is quantized in units of  $1/4$  i.e.  $J = -n/4$ . This is remarkable as it strongly suggests that if the usual spin-statistics connection is valid then whereas the Abelian charged vortex is an anyon with any phase factor, the non-Abelian ( $SU(2)$ ) charged vortex can only be a semion, a fermion or a boson.

#### 4.4 Relativistic Pure Chern-Simons Vortices

We have obtained above the charged vortex solutions in case the gauge part of the Abelian Higgs model consists of both the Maxwell and the CS term. It may be of some interest to enquire whether the Abelian Higgs model with pure CS term can also admit charged vortex solutions. This question is specially relevant in the context of condensed matter systems since in the long wave length limit, the CS term having one derivative dominates over the Maxwell term which has two derivatives. It turns out that the answer to the question is yes [37].

In the absence of the Maxwell term and with the same rotationally symmetric ansatz as in Eq. (45), it follows from Eqs. (48) and (49) that the gauge field equations are already of first order. However, Eq. (50), for the Higgs fields, is still a coupled second order equation. We now show that in case one replaces the standard double well  $\phi^4$ -type potential by the following  $\phi^6$ -type potential [38]

$$V(|\phi|) = \frac{e^4}{8\mu^2} |\phi|^2 (|\phi|^2 - C_0^2)^2 \quad (69)$$

then even the Higgs field satisfies a first order equation. It is worth pointing out here that whereas a Higgs potential of the type  $\sum_i C_i |\phi|^i$  with  $0 \leq i \leq 4$  is renormalizable in  $3+1$  dimensions,  $\sum_i C_i |\phi|^i$  with  $0 \leq i \leq 6$  is renormalizable in  $2+1$  dimensions.

When the Maxwell term is absent and the Higgs potential is as given by (69), the vortex energy (51) can be rewritten as

$$E_n = \pi C_0^2 \int_0^\infty r dr \left[ (f' \mp \frac{1}{r} f g)^2 + f^2 \left[ h \mp \frac{(1-f^2)}{2\delta} \right]^2 \mp \frac{1}{r} \frac{d}{dr} [(1-f^2)g] \right]. \quad (70)$$

This gives a rigorous lower bound on the energy in terms of the flux

$$E_n \geq \pm \pi C_0^2 [g(0) - g(\infty)] \equiv \pm \frac{1}{2} e C_0^2 \Phi \quad (71)$$

since the finite energy consideration requires that  $f^2 g$  vanish at both the ends. This bound is saturated when the following self-dual first order equations are satisfied

$$f'(r) = \pm \frac{1}{r} f g \quad (72)$$

$$-\frac{1}{r} g'(r) = \frac{h f^2}{\delta} = \pm \frac{1}{2\delta^2} f^2 (1 - f^2). \quad (73)$$

It is easily checked that these first order equations are consistent with the second order field Eq. (50). One can in fact decouple these coupled first order equations and show that the Higgs field  $f$  must satisfy the following un-coupled second order equation

$$f''(r) + \frac{1}{r} f'(r) - \frac{f'^2(r)}{f} + \frac{1}{2\delta^2} f^3 (1 - f^2) = 0. \quad (74)$$

Several comments are in order at this stage.



1. These self-dual equations are similar to those of the Nielsen-Olesen (neutral) self-dual vortices (which are valid only if  $\lambda = 1$ ).
2. Whereas the Lagrangian for the self-dual neutral vortex case (i.e. Lagrangian (43) with  $\mu = 0$  and  $\lambda = 1$ ) is the bosonic part of a  $N = 1$  supersymmetric theory [40], the Lagrangian for the self-dual charged vortex case (i.e. the Lagrangian (43) with the Maxwell term being absent and the Higgs potential being as given by Eq. (69)) is the bosonic part of a  $N = 2$  supersymmetric theory [41].
3. The  $\phi^4$ -potential as given in Eq. (43) and the  $\phi^6$ -potential as given by Eq. (69) represent very different physical situations. For example, whereas the  $\phi^4$ -potential corresponds to the case of the second order phase transition with  $T < T_c^{II}$ , the  $\phi^6$ -potential as given in Eq. (69) corresponds to the case of first order phase transition with  $T = T_c^I$  [42].
4. The nature of Higgs mechanism when only Chern-Simons term is present is somewhat unusual [39]. One finds that in the limit  $e^2 \rightarrow \infty$ ,  $\mu \rightarrow \infty$ , with their ratio fixed, the mass  $m_+$  decouples from the theory. Thus in the case of the pure CS term, one finds that after the Higgs mechanism, the gauge field is massive and propagates one mode.

Let us now discuss the most remarkable property of the self-dual Eqs. (72) and (73). In particular, since the Higgs potential (69) has degenerate minima at  $|\phi| = 0$  and  $|\phi| = C_0$ , hence, it turns out that at the self-dual point, one can simultaneously have both the topological and the non-topological charged vortex solutions. It is worth pointing out that at the time of this discovery, no other self-dual system was known which exhibited this remarkable property.

#### 4.5 Topological Self-dual Solutions

The topological, self dual charged vortex solutions satisfy the same boundary conditions as given by Eqs. (52) and (53) with  $\beta \equiv h(r=0) = \pm 1/2\delta^2$ . Note that the upper (lower) sign corresponds to  $n > 0 (< 0)$ . As a result, the flux  $\Phi$ , the Noether charge  $Q$ , and the angular momentum  $J$  of these charged vortices are again as given by Eqs. (54), (56) and (59) respectively while the energy of these charged vortices is  $\pi C_0^2 |n|$ . From now onwards, we shall confine our discussion to the case of  $n > 0$  i.e. those corresponding to the upper choice of sign. Solution with  $n < 0$  are related to these by the transformation  $g \rightarrow -g$ ,  $f \rightarrow f$ .

A countable infinite number of sum rules have been derived [43] and using the first two, it has been proved that the magnetic moment of the topological, self-dual charged  $n$ -vortex is given by [44]

$$K_z = 2\pi n(n+1) \frac{\delta^2}{e}. \quad (75)$$

Note that for the neutral  $n$ -vortex,  $K_z = \Phi = 2\pi n/e$ .

No analytic topological self dual charged vortex solution has been obtained as yet. However, one can show that all the fields approach their asymptotic values exponentially fast. It may be noted that at the self-dual point, the vector and the scalar meson masses are equal. Further, whereas for the Maxwell-CS case, the magnetic field is maximum at the core of the vortex ( $r \rightarrow 0$ ), for the pure CS vortices, the magnetic field is zero at the core of the vortex and is concentrated in a ring surrounding the vortex core.

#### 4.6 Non-topological Self-dual Solutions

Since  $|\phi| = 0$  as well as  $|\phi| = C_0$  are degenerate minima of the Higgs potential (69), hence it turns out that one could also have non-topological self-dual charged vortex solutions [44, 45]. In this case, the finite energy considerations demand the following boundary conditions

$$\lim_{r \rightarrow \infty} f(r) = 0, \quad g(r) = \mp \alpha, \quad \alpha > 0 \quad (76)$$

$$\lim_{r \rightarrow 0} f(r) = 0, \quad g(r) = n \quad \text{for } n \neq 0 \quad (77)$$

$$\lim_{r \rightarrow 0} f(r) = \eta, \quad g(r) = 0 \quad \text{for } n = 0 \quad (78)$$

where  $\eta$  is an arbitrary number while  $-\alpha(+\alpha)$  is for  $n0(<0)$ . As a result, the flux, the charge, the energy and the angular momentum of these vortices for  $(n0)$  are

$$\begin{aligned} \Phi &= \frac{2\pi}{e}(n + \alpha), \quad Q = \mu\Phi = \frac{2\pi\mu}{e}(n + \alpha), \\ J &= \frac{\pi\mu}{e^2}(\alpha^2 - n^2), \quad E = \pi C_0^2(n + \alpha). \end{aligned} \quad (79)$$

Note that unlike the topological case, the angular momentum is no more equal to  $-Q\phi/4\pi$ . Here  $\alpha$  is a positive number but how much is it? The finiteness of energy requires that  $\alpha > 1$  but otherwise  $\alpha$  seems to be completely arbitrary. However, it is not so and we now show [46] that  $\alpha$  satisfies a rigorous lower bound of  $\alpha \geq n + 2$ . To this end, consider the self-dual Eq. (73). On integrating both sides of this equation and using boundary conditions (76) to (78), one obtains (for  $n > 0$ )

$$-\int_0^\infty \frac{dg}{dr} dr = n + \alpha = \frac{1}{2\delta^2} \int_0^\infty r dr f^2 (1 - f^2) > 0. \quad (80)$$

Similarly, on using Eqs. (72) and (73) we have on integration

$$\int_0^\infty g \frac{dg}{dr} dr = \frac{1}{2}(\alpha^2 - n^2) = -\frac{1}{2\delta^2} \int_0^\infty r^2 f(1 - f^2) \frac{df}{dr} dr. \quad (81)$$

On integrating by parts and using the fact that  $r^2 f^2$  and  $r^2 f^4$  vanish as  $r \rightarrow \infty$  (note  $f(r) \sim r^{-\alpha}$  with  $\alpha > 1$  as  $r \rightarrow \infty$ ), we then have

$$(\alpha^2 - n^2) = \frac{1}{\delta^2} \int_0^\infty r dr (f^2 - \frac{1}{2} f^4). \quad (82)$$

On combining the two sum rules, we then have

$$(\alpha + n)(\alpha - n - 2) = \frac{1}{2\delta^2} \int_0^\infty r dr f^4 \geq 0 \quad (83)$$

which gives us a rigorous lower bound on  $\alpha$  i.e.  $\alpha \geq n + 2$ . It turns out that this bound is never saturated in the relativistic case. However, as we shall see below, it is indeed saturated in the case of the non-relativistic self-dual non-topological charged vortices. It may be noted here that there is however no upper bound on  $\alpha$ . We thus conclude that the flux of the relativistic non-topological vortices must necessarily be greater than  $4\pi(n + 1)/e$ . More remarkable is the fact that whereas the angular momentum of the topological vortices is always negative and proportional to  $n^2$ , the angular momentum of the non-topological vortices, on the other hand, is necessarily positive and in general is not proportional to  $n^2$ . Further, the magnetic moment of the non-topological vortices has also been computed analytically by using the sum rules and shown to be negative [46]

$$K_z = -\frac{2\pi\delta^2}{e}(\alpha + n)(\alpha - n - 1) < 0. \quad (84)$$

Note that the magnetic moment of the topological vortices is on the other hand always positive.

Are these non-topological vortices stable or do they decay to the charged scalar meson? This question has been discussed [47] and it has been shown that as far as the decay to the scalar meson is concerned, these non-topological solitons are at the edge of their stability. In particular, using  $E$  and  $Q$  as given by Eq. (79) and noting that the mass  $m$  of the scalar particle in the symmetric vacuum is  $e^2 c_0^2 / 2\mu$ , it follows that  $E = mQ/e$ . Thus the stability does not impose any upper bound on the charge of the non-topological soliton. No analytic solutions of Eqs. (72) and (73) have been obtained as yet in the non-topological self-dual case. However, the behavior of the fields near  $r \rightarrow 0$  and for large  $r$  is easily obtained. In particular, using the boundary conditions (76) to (78), it is not difficult to show that for  $r \rightarrow \infty$ , the  $n = 0$  vortex solution has the behavior

$$g(r) = -\alpha + \frac{G_0^2}{4(\alpha - 1)(r/\delta)^{2\alpha-2}} + O((r/\delta)^{-4\alpha+4}) \quad (85)$$

$$f(r) = \frac{G_0}{(r/\delta)^\alpha} - \frac{G_0^3}{8(\alpha-1)^2(r/\delta)^{3\alpha-2}} + O((r/\delta)^{-5\alpha+4}). \quad (86)$$

On the other hand, as  $r \rightarrow 0$ , while  $f(0)$  is not constrained,  $g(0)$  must vanish so as to have a non-singular solution. Thus for the  $n = 0$  non-topological vortex, the magnetic field  $(-g'(r)/r)$  is maximum at the core of the vortex ( $r = 0$ ) and falls off with a power law fall off as  $r \rightarrow \infty$ . Note, however, that the magnetic field for the topological CS vortices is zero at the core, and is maximum in a ring surrounding the core of the vortex.

Finally, let us consider the behavior of the  $n \neq 0$  (we as usual consider  $n > 0$ ) non-topological self-dual charged vortex solutions. It is easily shown that these solutions are hybrids of the two previous cases i.e. their large distance behavior is the same as those of the  $n = 0$  non-topological charged vortex solutions as given by Eqs. (85) and (86). On the other hand their short distance behavior is the same as those of the self-dual topological charged vortex solutions. Thus for  $n \neq 0$  non-topological vortices, the magnetic field vanishes at the core of the vortex and falls off with a power law fall off as  $r \rightarrow \infty$ .

It is worth pointing out that since the  $\phi^6$ -potential as given by Eq. (69) has two disconnected but degenerate vacua at  $|\phi| = 0$  and  $|\phi| = C_0$ , hence, apart from the charged vortex solutions, they also possess one dimensional domain wall solutions [45, 42].

So far, we have only discussed the self-dual rotationally symmetric CS vortices. However, the self-dual solutions can in fact be obtained even without choosing the rotationally symmetric  $n$ -vortex *ansatz* (45). Further, rigorous arguments have subsequently been given for the existence of the self-dual topological [48] and non-topological [49] charged vortex solutions even when the vortices are not superimposed on each other but lie at arbitrary positions in the plane. Let us note an interesting fact about the angular momentum of these charged vortices. For example, whereas the angular momentum of the  $n$  superimposed topological vortices is  $n^2$  times that of a single vortex, the angular momentum of the  $n$  topological vortices (each of which has unit vorticity) which are well separated from each other, is only  $n$  times the angular momentum of the single vortex. However, the energy, flux and the charge of the  $n$  vortices in both the cases is the *same*. Thus we see that whereas the energy, flux and charge, are the global quantities, the angular momentum of a configuration depends on the local behavior.

A zero-mode analysis of the spectrum of small fluctuations [45] around the self-dual vortices indicates that whereas the number of zero modes in the case of the topological self-dual vortices is  $2n$ , in the non-topological case, the same number is  $2n + 2[\alpha]$  where  $[\alpha]$  denotes the integer part of  $\alpha$ . In the topological case, this number is identified with the number of parameters required to describe the location of the  $n$  vortices while the counting is less clear in the non-topological case.

## 4.7 Interaction Between Self-Dual CS Vortices

The slow motion of the Abelian self-dual CS vortices has been analyzed [50] using Manton's technique [51]. In this approach, one constructs an effective quantum mechanical Lagrangian (not density) which describes the fluctuations about the static self-dual classical configurations and not surprisingly, one obtains a statistical interaction term. Further one also obtains a term corresponding to the velocity dependent Magnus force. It turns out that this force is in fact necessary in order to have correct spin-statistics relation.

Self-dual charged vortices have also been obtained in the original  $\phi^4$ -type model itself by adding a neutral scalar field to Eq. (43) and changing the  $\phi^4$ -potential suitably [53].

Finally, semi-local self-dual CS vortices have been obtained in an Abelian Higgs model with pure CS term [52] and with  $SU(N)_{global} \otimes U(1)_{local}$  symmetry. The interesting point is that the semi-local vortices, even though topologically trivial, are stable under small perturbations due to the gradient energy term.

#### 4.8 Non-relativistic Chern-Simons Vortices

Let us now discuss the non-relativistic limit of the Abelian Higgs model with the pure CS term. The Lagrangian density for the Abelian Higgs model with pure CS term is given by

$$\mathcal{L} = \frac{1}{2}(D_\mu \phi)^*(D^\mu \phi) + \frac{\mu}{4}\epsilon_{\mu\nu\lambda}F^{\mu\nu}A^\lambda - \frac{e^4}{8c^4\mu^2}|\phi|^2(|\phi|^2 - C_0^2)^2, \quad (87)$$

where the Higgs potential is as given by Eq. (69). Here we write all the factors of the velocity of light  $c$  explicitly since we are considering the non-relativistic limit of a relativistic theory. Let us first note that the quadratic term in the Higgs potential defines the mass through its coefficient  $m^2c^2/2$ . Comparison with Eq. (87) shows that  $C_0^2$  must have the value  $C_0^2 = (2|\mu|mc^3)/e^2$  so that the Lagrangian density (87) can be rewritten as

$$\begin{aligned} \mathcal{L} = & \frac{1}{2c^2}|\partial_t - \frac{ie}{\hbar}A^0\phi|^2 - \frac{1}{2}|\mathbf{D}\phi|^2 - \frac{m^2c^2}{2}|\phi|^2 \\ & + \frac{me^2}{2c|\mu|}|\phi|^4 - \frac{e^4}{8c^4\mu^2}|\phi|^6 + \frac{\mu}{4}\epsilon_{\mu\nu\lambda}F^{\mu\nu}A^\lambda. \end{aligned} \quad (88)$$

The non-relativistic limit ( $c \rightarrow \infty$ ) now proceeds in the standard manner. On writing the mode expansion of the scalar field  $\phi$  as

$$\phi = \frac{1}{\sqrt{m}} \left[ e^{-imc^2t} \psi + e^{imc^2t} \bar{\psi}^* \right] \quad (89)$$

and substituting it in Eq. (88), dropping all terms that either oscillate as  $c \rightarrow \infty$  or are sub-leading in powers of  $c$ , the matter part of the Lagrangian density can be shown to be

$$\mathcal{L} = i\psi^*D_0\psi - \frac{1}{2m}|\mathbf{D}\psi|^2 + \frac{e^2}{2mc|\mu|}\rho^2 + \frac{\mu}{4}\epsilon_{\mu\nu\lambda}F^{\mu\nu}A^\lambda. \quad (90)$$

Here  $\rho = \psi^*\psi$  is the matter density of particles and we have dropped the anti-particle part from the Lagrangian density (i.e. we are working in the zero anti-particle sector) by setting  $\bar{\psi} = 0$  since the particle and the anti-particle parts are separately conserved. The remarkable fact is that one now has an *attractive* quartic ( $\rho^2$ ) self-interaction. This non-relativistic model can be looked upon either as a non-relativistic classical field theory or as a second quantized  $N$ -body problem with 2-body attractive delta-function interaction.

The Euler-Lagrangian equations of motion which follow from the Lagrangian density (90) are

$$-\frac{1}{2m}\mathbf{D}^2\psi - \frac{e^2}{mc|\mu|}|\psi|^2\psi - iD_0\psi = 0 \quad (91)$$

$$F_{\mu\nu} = -\frac{1}{\mu}\epsilon_{\mu\nu\rho}J^\rho \quad (92)$$

where  $J^\mu \equiv (\rho, \vec{J})$  is a Lorentz covariant notation for the conserved non-relativistic charge and current densities i.e.

$$\rho = |\psi|^2, \quad J^k = -\frac{i\hbar^2}{2m}[\psi^*D^k\psi - (D^k\psi)^*\psi]. \quad (93)$$

The field Eqs. (91) and (92) are together termed as the *planar gauged nonlinear Schrödinger equations*. The gauge field Eq. (92) can also be re-expressed as

$$B \equiv F_{12} = \frac{e}{\mu}\rho \quad (94)$$

$$E^i \equiv F_{i0} = -\frac{e}{c\mu}\epsilon^{ik}J_k. \quad (95)$$

From here, we immediately obtain the fundamental relation between the Noether charge  $Q$  and the magnetic flux  $\Phi$  i.e.  $Q = \mu\Phi$ . As in the relativistic case, it is easily checked that the second order field Eqs. (91) and (92) are solved by Eq. (94) and the self-dual ansatz

$$D_j\psi = \pm i\varepsilon_{jk}D_k\psi \quad (96)$$

in the case of the static solutions with  $A_0$  chosen as

$$A_0 = \mp \frac{e}{2m\mu c} |\psi|^2. \quad (97)$$

Here we have made use of the following factorization identity

$$D^2\psi = D_{\pm}D_{\mp}\psi \mp \frac{e}{c}F_{12}\psi. \quad (98)$$

We now show that the self-dual Eqs. (94) and (96) can be solved completely and explicitly. On writing the complex field  $\psi$  as  $\psi = e^{-i\omega} \rho^{1/2}$  the self-duality Eq. (96) yields the vector potential

$$A_i = \partial_i\omega \pm \frac{c}{2e}\varepsilon^{ij}\partial_j \ln \rho \quad (99)$$

which is valid away from the zeros of  $\rho$ . On inserting this form of  $A$  into the other self-dual Eq. (94) yields the famous Liouville equation

$$\nabla^2 \ln \rho = -\frac{2e^2}{c|\mu|}\rho \quad (100)$$

which is known to be integrable and completely solvable and which must be solved away from the zeros of  $\rho$ . It is worth noting that with our sign conventions, we have the Liouville equation with the correct sign in that only such an equation has real, positive, regular solutions. The most general such solution is known to be given by

$$\rho = \frac{c|\mu||f'(z)|^2}{e^2[1+|f(z)|^2]} \quad (101)$$

where  $f(z)$  is any holomorphic function and  $z = re^{i\theta}$ . Explicit radially symmetric solutions may be obtained by taking  $f(z) = (z/z_0)^{\pm n}$ . The corresponding self-dual charge density is

$$\rho = \frac{4|\mu|n^2c}{e^2r_0^2} \frac{(r/r_0)^{2(n-1)}}{[1+(r/r_0)^{2n}]^2} \quad (102)$$

which behaves like  $r^{2(n-1)}$  as  $r \rightarrow 0$  while as  $r \rightarrow \infty$ , it behaves like  $r^{-2-2n}$ . Thus  $\rho$  is regular at the origin if  $n \geq 1$ . From Eq. (99) it then follows that as  $r \rightarrow 0$ , the vector potential behaves as

$$A_i(r) \sim \partial_i\omega \pm \frac{c(n-1)}{e}\varepsilon_{ij}\frac{x^j}{r^2} \quad (103)$$

i.e. it is singular at  $r = 0$ . This singularity is removed if we choose  $\omega = \pm c(n-1)\theta/e$ . Thus the profile of the self-dual  $\psi$  field is given by

$$\psi(r) = \frac{2n\sqrt{|\mu|c}}{er_0} \frac{(r/r_0)^{n-1}}{[1+(r/r_0)^{2n}]} e^{\pm i(n-1)\theta}. \quad (104)$$

On requiring that  $\psi$  be single valued, we then find that  $n$  must be an integer, and for  $\rho$  to have decaying behavior as  $r \rightarrow \infty$ , we require that  $n$  must be positive.

Several comments are in order at this stage.

1. Integrating  $\rho$  as given in (102) over all space yields  $n$  (the total number of particles) and hence the flux (in view of Eq. (94)). We obtain  $\Phi = (4\pi cn/e)$  with  $n = 1, 2, \dots$  which means that this configuration carries an even number of flux units. This is in contrast to the relativistic case where the flux unit need not necessarily be even. Further, note that unlike the relativistic non-topological case, here the lower bound on  $\alpha (\geq n+2)$  is saturated. As has been shown [55], this is because of the special inversion symmetry of the Liouville equation. In particular, notice that the Liouville equation is invariant under the transformations

$$r \rightarrow 1/r, \theta \rightarrow \theta, \rho(r) \rightarrow \rho(1/r) = r^4 \rho(r). \quad (105)$$

As a result, the behavior of  $\rho$  at infinity is uniquely determined by its behavior at the origin thereby fixing  $\alpha = n + 2$ .

2. It is worth pointing out the  $Q, \Phi$  and  $J$  for the non-relativistic charged vortices are the same as those for the relativistic non-topological charged vortices as given by Eq. (79) provided one chooses  $\alpha = n + 2$  (note that in the non-relativistic case,  $n = 1, 2, \dots$  while  $n = 0, 1, 2, \dots$  in the relativistic case).
3. The radially symmetric solution (104) was obtained by choosing the holomorphic function  $f(z) \propto (z)^{-n}$  and corresponds to  $n$  solitons superimposed at the origin with common scale factor  $r_0$ . The most general solution corresponding to  $n$  separated solitons may be obtained by taking

$$f(z) = \sum_{i=1}^n \frac{\alpha_i}{(z - z_i)} \quad (106)$$

where  $2n$  real parameters  $z_i$  describe the location of the solitons and  $2n$  real parameters  $\alpha_i$  correspond to the scales and the phases of the solitons. Thus the solution depends on  $4n$  parameters. Using an index theory calculation [56] it has been shown that this is the most general solution.

## 5 $CP^1$ Solitons With Hopf Term

In this section we discuss the extended (neutral) anyon solutions in relativistic field theories. Historically, such solutions were first written down in the case of  $O(3)$   $\sigma$ -model with Hopf term in 2+1 dimensions [57]. Unfortunately, in this case, the Hopf term cannot be written down as a local function of the basic fields of the theory. Therefore, we shall discuss the essentially equivalent example of the  $CP^1$  model with the Hopf term since in this case the Hopf term can be written down as a local function of the basic fields of the theory [58].

The action for the  $CP^1$  model in 2+1 dimensions is given by

$$I = \int d^3x (D_\mu z)^* (D^\mu z) \quad (107)$$

where  $D_\mu z \equiv (\partial_\mu - iA_\mu)z$  with  $z = (z_1, z_2)$  being a complex vector fulfilling  $|z|^2 = 1$ . Note that  $A_\mu$  here does not represent independent degrees of freedom, but is entirely determined in terms of  $z(x)$  through the constraint equation

$$A_\mu = -iz^* \partial_\mu z. \quad (108)$$

The action (107) is invariant under the local  $U(1)$  transformations

$$z_a(x) \rightarrow z_a(x) e^{i\Lambda(x)}, \quad A_\mu(x) \rightarrow A_\mu(x) + \partial_\mu \Lambda(x). \quad (109)$$

As is well known, the  $CP^1$  model admits self-dual soliton solutions. To obtain them, let us first note that the field equation is obtained by extremizing the action (107) with respect to  $z(x)$  subject to the constraint  $|z|^2 = 1$ . This constraint is best introduced in the variational formalism

by using a Lagrangian multiplier i.e. one extremizes  $I + \int d^3x \lambda(x)(z^* z - 1)$ . The resulting field equation is

$$(D_\mu D^\mu + \lambda)z = 0. \quad (110)$$

The Lagrange multiplier  $\lambda(x)$  is eliminated by using  $\lambda = \lambda z^* z = -z^* D_\mu D^\mu z$ . Let us now consider the static solutions. In this case, the field equation (110) reduces to

$$\nabla^2 z - (z^* \cdot \nabla^2 z)z = 0. \quad (111)$$

The energy of a static solution as obtained from the action (107) is clearly

$$E = \int (D_i z)^* (D_i z) d^2x, \quad i = 1, 2. \quad (112)$$

Finiteness of energy requires that as  $r \equiv |x| \rightarrow \infty$ ,  $D_i z \equiv \partial_i z - i A_i z = 0$ .

Let us start from the topological inequality which follows from

$$\left[ (D_i z)^* \pm i \varepsilon_{ij} (D_j z)^* \right] \cdot \left[ D_i z \mp i \varepsilon_{ik} D_k z \right] \geq 0. \quad (113)$$

Because of the constraint  $|z|^2 = 1$ , this inequality can be re-expressed in the form

$$(D_i z)^* \cdot (D_i z) \geq \varepsilon_{ij} (D_i z)^* \cdot (D_j z) \quad (114)$$

so that the energy is bounded from below by the topological charge  $Q$  i.e.  $E \geq 2\pi |Q|$ , where

$$Q = -\frac{i}{2\pi} \int d^2x \varepsilon_{ij} (D_i z)^* \cdot (D_j z). \quad (115)$$

In any  $Q$ -sector, the energy reaches its minimum when the fields minimize the energy in that sector and satisfy the first order self dual field equation

$$D_i z = \pm i \varepsilon_{ij} D_j z. \quad (116)$$

Note that the solutions of Eq. (116) automatically solve the second order field Eq. (111) while the converse need not be true.

The most general solution for  $z$  can be written down in terms of (anti) holomorphic function  $\omega$

$$z = \frac{1}{\sqrt{1+|\omega|^2}} \begin{pmatrix} \omega \\ 1 \end{pmatrix}. \quad (117)$$

These solutions are characterized by the energy  $E = 2\pi |Q|$  where  $Q$  is as given by Eq. (115). One can in fact define a topological current  $J^\mu$

$$J^\mu = -\frac{i}{2\pi} \varepsilon^{\mu\nu\lambda} (D_\nu z)^* (D_\lambda z) \quad (118)$$

which is conserved by construction, and the topological charge  $Q$  as given above, is related to it by  $Q = \int J^0 d^2x$ . One can easily show that for the soliton solutions,  $Q$  is just the winding number i.e.  $Q$  clearly describes the homotopy of the mapping  $S_2 \rightarrow S_2$ .

Since  $J^\mu$  is the topological conserved current, hence one can clearly add the following gauge invariant action

$$I_H = \int d^3x \frac{\theta}{2\pi} A_\mu J^\mu \quad (119)$$

to the original action (107). This action is nothing but the Hopf term which is related formally to the CS term since from Eqs. (108) and (118) it follows that

$$A_\mu J^\mu = \frac{1}{4\pi} \varepsilon^{\mu\nu\lambda} A_\mu F_{\nu\lambda}. \quad (120)$$

Note however that here  $A_\mu$  is not an independent gauge field but is entirely determined in terms of  $z(x)$  through the constraint Eq. (108). As a result, unlike the CS term, the Hopf term is locally a total divergence and hence does not contribute to the equations of motion.

Note that unlike the CS term, the Hopf term has no dynamics. Besides, for the  $CP^1$  soliton solutions (which are time independent solutions of the equations of motion), the Hopf term is identically zero because of the time derivative and the relationship (108). Thus the way the Hopf term imparts fractional spin and statistics to the soliton is similar to that in quantum mechanics but it is very different than the way the CS term imparts fractional spin and statistics. In particular, since the Hopf density is a total divergence, hence the Hopf action can be expressed in terms of the surface terms, namely two integrals at the initial and final times so that in the path integral formalism, the contribution of this action is essentially in terms of the phases of the initial and the final wave functions. Since the configuration space in question is multi-connected, the Hopf action depends on the homotopy classes of the path and, therefore, the converted phases are multi-valued which in turn gives rise to the fractional spin ( $= \theta/2\pi$ ) and the solitons obey fractional statistics characterized by  $\theta$  [57, 58].

## 6 Anyons as Elementary field Quanta

In this section we enquire whether one can construct local quantum field theories where the fundamental fields represent the creation and annihilation of anyons. Let us consider a complex bosonic non-relativistic matter field  $\psi(\mathbf{x}, t)$  of mass  $m$  (of course a similar discussion can also be done for the fermionic matter field). Let us minimally couple it to an Abelian gauge field  $A_\mu$  with a CS kinetic term [5, 59]

$$S = \int d^3x [i\psi^\dagger D_0\psi + \frac{1}{2m}\psi^\dagger (D_1^2 + D_2^2)\psi + \frac{\mu}{2}\epsilon^{\mu\nu\lambda}A_\mu\partial_\nu A_\lambda] \quad (121)$$

where  $D_\mu = \partial_\mu - ieA_\mu$  is the covariant derivative. For simplicity, in this section we shall set  $\hbar = c = 1$ . On varying the action with respect to  $A_\mu$ , we obtain

$$\epsilon^{\mu\nu\lambda}F_{\nu\lambda} = \frac{2e}{\mu}J^\mu \quad (122)$$

where the current  $J^\mu$  is explicitly given by

$$\rho \equiv J^0 = \psi^\dagger\psi, \quad J^k = \frac{1}{2mi}[\psi^\dagger D^k\psi - (D^k\psi)^\dagger\psi]. \quad (123)$$

Here  $\rho$  and  $\mathbf{J}$  are the number density and the current density operators respectively which satisfy the continuity equation  $\partial_t\rho + \nabla\cdot\mathbf{J} = 0$ . As seen in previous sections, Eq. (122) is a remarkable relation indicating that the CS field strength is completely determined by the particle current. Even more remarkable is the fact that the gauge potential  $A_\mu$  itself is not an independent degree of freedom.

Let us consider the  $\mu = 0$  component of Eq. (122)

$$B = \frac{e}{\mu}\rho \quad (124)$$

where  $B = \nabla \times \mathbf{A}$  is the CS magnetic field. This equation is clearly the second quantized version of the Gauss law constraint obtained in the last two chapters (except that whereas in those cases  $\rho$  was the charge density, here  $\rho$  is the matter density, hence the extra factor of  $e$  in Eq. (124) compared to those cases). Now, in the weyl gauge  $\partial_i A^i = 0$ . Hence, one can invert Eq. (124) without any ambiguity and solve for the vector potential  $\mathbf{A}$ . We obtain

$$A^i(x) = \epsilon^{ij} \frac{\partial}{\partial x^j} \left( \frac{e}{\mu} \int d^2y G(\mathbf{x} - \mathbf{y}) \rho(y) \right) \quad (125)$$



where  $G$  is the two-dimensional Green function

$$\nabla^2 G(\mathbf{x} - \mathbf{y}) = \delta(\mathbf{x} - \mathbf{y}) \quad (126)$$

whose solution is well known to be

$$G(\mathbf{x} - \mathbf{y}) = \frac{1}{2\pi} \ln(p |\mathbf{x} - \mathbf{y}|) \quad (127)$$

where  $p$  is an arbitrary scale. Thus  $A^i$  can be written as

$$\begin{aligned} A^i(x) &= \varepsilon^{ij} \frac{\partial}{\partial x^j} \left[ \frac{e}{2\pi\mu} \int d^2y \ln |\mathbf{x} - \mathbf{y}| \rho(y) \right] \\ &= -\frac{e}{2\pi\mu} \int d^2y \frac{\partial}{\partial x^i} \phi(\mathbf{x} - \mathbf{y}) \rho(y) \end{aligned} \quad (128)$$

where  $\phi$  is the winding (polar) angle i.e.

$$\phi(\mathbf{x} - \mathbf{y}) = \arctan\left(\frac{x^2 - y^2}{x^1 - y^1}\right). \quad (129)$$

Note that while writing the second line of Eq. (128), we have used the Cauchy-Riemann equations

$$\varepsilon^{ij} \frac{\partial}{\partial x^j} \ln |\mathbf{x} - \mathbf{y}| = -\frac{\partial}{\partial x^i} \phi(\mathbf{x} - \mathbf{y}). \quad (130)$$

It is worth pointing out that  $\varepsilon^{ij} \frac{\partial}{\partial x^j} G(\mathbf{x} - \mathbf{y})$  is ill-defined at  $\mathbf{x} = \mathbf{y}$ . Thus one has to supplement Eqs. (126) and (127) with a regularization prescription. One such prescription is

$$\varepsilon^{ij} \frac{\partial}{\partial x^j} G(\mathbf{x}) \longrightarrow \varepsilon^{ij} \frac{\partial}{\partial x^j} G^a(\mathbf{x}) \quad (131)$$

where the regulated Green function  $G^{(a)}(\mathbf{x})$  is

$$G^{(a)}(\mathbf{x}) = \frac{1}{a\pi} \int d^2y \left( \frac{1}{2\pi} \ln |\mathbf{x} - \mathbf{y}| \right) e^{-y^2/a}. \quad (132)$$

This has the desired property that

$$\lim_{a \rightarrow 0} G^{(a)}(\mathbf{x}) = G(\mathbf{x}) = \frac{1}{2\pi} \ln |\mathbf{x}| \quad (133)$$

while for any  $a$

$$\lim_{x \rightarrow 0} \varepsilon^{ij} \frac{\partial}{\partial x^j} G^{(a)}(\mathbf{x}) = 0 \quad (134)$$

so that once Eq. (132) is systematically used, all ambiguities are eliminated.

If one is now allowed to move the derivative operator outside the integral (128), then one could express  $\mathbf{A}$  as a gradient. However,  $\phi(\mathbf{x} - \mathbf{y})$  is a multi valued function. Hence one must first fix a branch-cut in the  $y$ -plane starting at  $x$  so as to make it single-valued. No matter what choice is made for this cut, the resulting range of integration of  $\mathbf{y}$  will depend on  $\mathbf{x}$  and hence extra contributions are produced in moving  $\partial/\partial x^i$  outside the  $\mathbf{y}$  integral. Thus, in general one can not write

$$\mathbf{A}(\mathbf{x}) = -\frac{e}{2\pi\mu} \nabla_x \left[ \int d^2y \phi(\mathbf{x} - \mathbf{y}) \rho(y) \right]. \quad (135)$$

so that in general  $\mathbf{A}$  is not a pure gauge and hence it cannot be removed by a gauge transformation. However, in the special case when  $\rho(y)$  is a sum of  $\delta$ -functions,  $\mathbf{A}(x)$  is indeed a pure gauge. Such a situation arises in the case of non-relativistic localized point particles [59]. Let us assume that in the context of our non-relativistic model (121) too,  $\rho(y)$  is a sum of  $\delta$ -functions in which case the CS gauge field  $A_\mu$  is entirely determined by the matter configuration i.e.  $\rho$  and  $\mathbf{J}$ .

Thus, in the case of localized densities,  $A_\mu(x) = -\partial_\mu \Lambda(x)$  i.e. the CS field is a pure gauge and hence it can be removed by the gauge transformation  $A_\mu \rightarrow A'_\mu = A_\mu + \partial_\mu \Lambda = 0$ . Thus, under such a singular transformation, covariant derivatives turn into ordinary derivatives, and the action (121) becomes

$$S' = \int d^3x \left[ i\tilde{\psi}^+ \partial_0 \tilde{\psi} + \frac{1}{2m} \tilde{\psi}^+ (\partial_1^2 + \partial_2^2) \tilde{\psi} \right] \quad (136)$$

where the new matter field  $\tilde{\psi}$  is defined as

$$\tilde{\psi}(x) = e^{-ie\Lambda(x)} \psi(x), \quad \tilde{\psi}^+(x) = \psi^+(x) e^{ie\Lambda(x)}. \quad (137)$$

The above action (136) is that of a free, complex, non-relativistic, scalar field  $\tilde{\psi}$ . However, we now show that such a field *does not* obey the conventional commutation relations as satisfied by  $\psi$ .

We can quantize the action (121) by imposing the equal-time commutation relations for the bosonic field  $\psi$

$$[\psi(\mathbf{x}, t), \psi^+(\mathbf{y}, t)] = \delta(\mathbf{x} - \mathbf{y}) \quad (138)$$

$$[\psi(\mathbf{x}, t), \psi(\mathbf{y}, t)] = 0 = [\psi^+(\mathbf{x}, t), \psi^+(\mathbf{y}, t)]. \quad (139)$$

Since the gauge field  $\mathbf{A}$  is a function of the number density operator  $\rho (= \psi^+ \psi)$ , hence the commutator of  $\mathbf{A}$  and  $\psi$  is not trivial. In fact using Eqs. (128) and (138) we obtain

$$[A^i(\mathbf{x}, t), \psi(\mathbf{y}, t)] = -\frac{e}{\mu} \epsilon^{ij} \frac{\partial}{\partial x^j} G(\mathbf{x} - \mathbf{y}) \psi(\mathbf{y}). \quad (140)$$

On using the regularized Green function as given by Eq. (132), it then follows by using Eq. (134) that  $[A^i(\mathbf{x}, t), \psi(\mathbf{x}, t)] = 0$ . This is interesting because it means that there are no ordering ambiguities in the quantum theory as given by Eq. (121).

One can now show that when  $\psi$  obeys ordinary commutation relations,  $\tilde{\psi}$  obeys

$$\tilde{\psi}(\mathbf{x}, t) \tilde{\psi}(\mathbf{y}, t) = e^{i\pi\alpha} \tilde{\psi}(\mathbf{y}, t) \tilde{\psi}(\mathbf{x}, t) \quad (141)$$

i.e. the matter field  $\tilde{\psi}$  obeys anyonic commutation relations of statistics  $\alpha (= e^2/2\pi\mu)$ . If instead, we make a cut along the negative  $x'$ -axis, then we would obtain a phase factor ( $e^{-i\pi\alpha}$ ), opposite to that in Eq. (141). Proceeding in the same way, it is easily shown that if  $\mathbf{x} \neq \mathbf{y}$  then

$$\tilde{\psi}(\mathbf{x}, t) \tilde{\psi}^+(\mathbf{y}, t) = e^{-i\pi\alpha} \tilde{\psi}^+(\mathbf{y}, t) \tilde{\psi}(\mathbf{x}, t). \quad (142)$$

It must however be noted that for  $\mathbf{x} = \mathbf{y}$ , the phase proportional to  $\alpha$  vanishes and hence the canonical commutation relations remain unchanged.

Some clarification is called for at this stage. What one has shown is that the fields  $\tilde{\psi}(\mathbf{x}, t)$ ,  $\tilde{\psi}(\mathbf{y}, t)$  satisfy anyonic commutation relations with the phase factor  $e^{+i\pi\alpha}$  or  $e^{-i\pi\alpha}$  depending on how we make the cut. However, this is not enough. What is really required is that the phase of the wave function changes both by  $+\pi\alpha$  and  $-\pi\alpha$  in response to which way we braid in interchanging  $\mathbf{x}$  and  $\mathbf{y}$ . No one has been able to show this so far. In fact, what we have shown above is the best that one can achieve with local operators  $\tilde{\psi}$ ,  $\tilde{\psi}^+$ . Local information, like initial and final positions of particles, is simply not sufficient to code the braiding, where we also have to specify which way the particles passed around each other in interchanging their positions. As I see it, the only way to take care of this problem in this formalism is to choose such a definition of the multi-valued function  $\phi$  which will make  $\tilde{\psi}$  a non-local operator.

Summarizing, it appears that within the non-relativistic field theory formalism, anyons can only be described by non-local operators, which are hard to deal with. If one insists on a local formulation, then one has to hide the statistics in an interaction with a CS field.

There is no doubt that ideally the various effects of fractional spin, such as the spin-statistics theorem should be understood only in a full fledged relativistic quantum field theory. However, relatively little is known in this respect. The point is, if the fundamental fields are to carry fractional spin, they must carry a multi-valued irreducible representation of  $SO(2,1)$ . This is

because, a rotation of  $2\pi$  does not leave the Wave function invariant, but rather, it multiplies it by a phase  $e^{2i\pi j}$ . We then have the following two options.

The first option is that we define infinite component fields and from them construct one particle dynamics by imposing equations of motion that satisfy the requirement that one-particle states provide multi-valued Poincaré equations. The most difficult part is the derivation of an action that reproduces these equations of motion. This requires handling a nonlocal theory and no one really knows how to quantize such a theory.

The second option is to work with multi-valued fields by adding the CS term to the action and essentially repeat what we have done above for the non-relativistic case. Thus, instead of the non-relativistic model (121), one could consider a relativistic field theory, say a complex scalar field theory, coupled to an Abelian gauge field with a CS kinetic energy term (and no Maxwell term). Coming back to complex fields, one again wants to know if one can construct local quantum field theory where the fundamental fields represent the creation and annihilation of anyons. On proceeding exactly as in the nonrelativistic case, one again obtains Eq. (124). However, now the particles are not point particles but are extended objects, hence  $\rho(y)$  cannot be a sum of delta functions. Thus it is not possible to write  $\mathbf{A}$  as a pure gauge and hence it cannot be removed by a gauge transformation. Thus, it is not at all clear whether in the relativistic case the only effect of the gauge field is to endow the particle with arbitrary spin or if residual interactions are also present. A similar problem also arises in models which emerge from the relativistic theory in the non-relativistic limit. In particular, one obtains different results depending on which limit is taken first i.e. the size of the extended object going to zero vis-a-vis the regulator parameter going to zero. Attempts have been made to tackle these problems by quantizing the theory with CS term on a lattice with or without the Maxwell term. So far, these attempts have met with only a limited success.

Thus it is fair to say that, so far we do not have a model in relativistic local quantum field theory where the fundamental (non-interacting) field quanta are themselves anyons. In fact it appears unlikely that one can obtain a simple, local (relativistic) Lagrangian for anyons. This is because, even in  $2 + 1$  dimensions, spin has to be an integer or half-integer for local fields. On the other hand, fractional spin is admissible for fields which carry charges associated with gauge symmetries (with accompanying flux integrals at infinity) which are typically localizable only in space-like cones [32, 30]. This is what happens for example, when one generates fractional spin by coupling point particles to a CS gauge field which has non-trivial long-ranged properties.

## References

- [1] E.A. Abbot, *Flatland* (Princeton Univ. Press, New Edition, 1991).
- [2] C.H. Hinton, *An Episode Of Flatland* (1907).
- [3] M. Gardner, in *The Unexpected Hanging And Other Mathematical Diversions*, ed. M. Gardner (Simon and Schuster 1969).
- [4] A.K. Dewdney, *Two Dimensional Science and Technology*, *J. Recreate. Math.* **12** (1979) 16 ; For a short summary of the book see M. Gardner, *Sci. Ame.*, *July Issue* (1980) 18.
- [5] For the details see for example, A. Khare, *Fractional Statistics and Quantum Theory* (World Scientific, Singapore, 1997).
- [6] R.B. Laughlin, *Phys. Rev. Lett.* **50** (1983) 1395.
- [7] For a popular readable account see A. Khurana, *Phys. Today* **43**(1) (1990) 19.
- [8] J.M. Leinaas and J. Myrheim, *Nuovo Cim.* **B37** (1977) 1.
- [9] R. Mirman, *Nuovo Cim.* **B18** (1973) 110.
- [10] F. Haldane, *Phys. Rev. Lett.* **67** (1991) 937.

- [11] For a detailed review of various properties see for example, A. Khare, *Fort. der. Physik* **38** (1990) 507 ; *Proc. Indian Natn. Sc. Acad.* **A61** (1995) 161.
- [12] W. Siegel, *Nucl. Phys.* **B156** (1979) 135; J. Schonfeld, *Nucl. Phys.* **B185** (1981) 157; R. Jackiw and S. Templeton, *Phys. Rev.* **D23** (1981) 2291; C.R. Hagen, *Ann. Phys.N.Y.* **157** (1984) 342.
- [13] S. Deser, R. Jackiw and S. Templeton, *Phys. Rev. Lett.* **48** (1982) 975; *Ann. of Phys.* **140** (1982) 372.
- [14] S. Coleman and B. Hill, *Phys. Lett.* **B159** (1985) 184.
- [15] A. Khare, R.B. MacKenzie and M.B. Paranjape, *Phys. Lett.* **B343** (1995) 239.
- [16] C. R. Hagen, P.K. Panigrahi and S. Ramaswami, *Phys. Rev. Lett.* **61** (1988) 389.
- [17] L.D. Landau and E.M. Lifshitz, *Electrodynamics of Continuous Media, Second Edition* (Pergamon Press, Oxford 1963).
- [18] T.H. O'Dell, *The Electrodynamics of Magneto-Electric Media* (North-Holland, Amsterdam, 1970).
- [19] A. Khare and T. Pradhan, *Phys. Lett.* **B231** (1989) 178.
- [20] S.K. Paul and A. Khare, *Phys. Lett.* **B193** (1987) 253, **B196** (1987) E571.
- [21] R.D. Pisarski and S. Rao, *Phys. Rev.* **D32** (1985) 2081.
- [22] G. Giavarini, C.P. Martin and F. Ruiz Ruiz, *Nucl. Phys.* **B381** (1992) 222.
- [23] A. Khare, R.B. MacKenzie, P.K. Panigrahi and M.B. Paranjape, *Phys. Lett.* **B355** (1995) 236; L. Chen, G. Dunne, K. Haller and E. Lim-Lombridas, *Phys. Lett.* **B348** (1995) 468.
- [24] A.N. Redlich, *Phys. Rev. Lett.* **52** (1984) 18 ; *Phys. Rev.* **D29** (1984) 2366; A. J. Niemi and G.W. Semenoff, *Phys. Rev. Lett.* **51** (1983) 2077.
- [25] A.S. Schwarz, *Lett. Math. Phys.* **2** (1978) 247; E. Witten, *Comm. Math. Phys.* **121** (1989) 351.
- [26] A.A. Abrikosov, *Sovt. Phys. JETP* **5** (1957) 1174.
- [27] H.B. Nielsen and P. Olesen, *Nucl. Phys.* **B61** (1973) 45.
- [28] B. Julia and A. Zee, *Phys. Rev.* **D11** (1975) 2227.
- [29] S.K. Paul and A. Khare, *Phys. Lett.* **B174** (1986) 420; **B177** (1986) E453.
- [30] J. Fröhlich and P.A. Marchetti, *Comm. Math.-Phys.* **121** (1989) 177.
- [31] E. Witten, *Phys. Lett.* **B86** (1979) 283.
- [32] D. Buchholz and K. Fredenhagen, *Comm. Math. Phys.* **84** (1982) 1.
- [33]
- [34] L. Jacobs, A. Khare, C.N. Kumar and S.K. Paul, *Int. J. Mod. Phys.* **A6** (1991) 3441.
- [35] S.K. Paul and A. Khare, *Phys. Lett.* **B171** (1986) 244.
- [36] H.J. de Vega and F.A. Schaposnik, *Phys. Rev. Lett.* **56** (1986) 2564 ; *Phys. Rev.* **D34** (1986) 3206; C.N. Kumar and A. Khare, *Phys. Lett.* **B178** (1986) 395, **B182** (1986) E415; *Phys. Rev. Lett.* **59** (1987) 377 ; *Phys. Rev.* **D36** (1987) 3253.

- [37] D.P. Jatkar and A. Khare, *Phys. Lett.* **B236** (1990) 283.
- [38] J. Hong, Y. Kim and P.Y. Pac, *Phys. Rev. Lett.* **64** (1990) 2230; R. Jackiw and E.J. Weinberg, *Phys. Rev. Lett.* **64** (1990) 2234.
- [39] S. Deser and Z. Yang, *Mod. Phys. Lett.* **A4** (1989) 2123.
- [40] P. di Vecchia and S. Ferrara, *Nucl. Phys.* **B130** (1977) 93.
- [41] C. Lee, K. Lee and E.J. Weinberg, *Phys. Lett.* **B243** (1990) 105.
- [42] S.N. Behera and A. Khare, *Pramana (J. Phys., India)* **15** (1980) 245.
- [43] A. Khare, *Phys. Lett.* **B277** (1992) 123.
- [44] A. Khare, *Phys. Lett.* **B255** (1991) 393.
- [45] R. Jackiw, K. Lee and E.J. Weinberg, *Phys. Rev.* **D42** (1990) 3488.
- [46] A. Khare, *Phys. Lett.* **B263** (1991) 227.
- [47] D.P. Jatkar and A. Khare, *J. Phys.* **A24** (1991) L1201; D. Bazeia, *Phys. Rev.* **D43** (1991) 4074.
- [48] R. Wang, *Comm. Math. Phys.* **137** (1991) 587.
- [49] J. Spruck and Y. Yang, *Comm. Math. Phys.* **149** (1992) 361.
- [50] S.K. Kim and H. Min, *Phys. Lett.* **B281** (1992) 81.
- [51] N. Manton, *Phys. Lett.* **B110** (1982) 54 ; *Phys. Lett.* **B154** (1985) 397.
- [52] A. Khare, *Phys. Rev.* **D46** (1992) R 2287.
- [53] C. Lee, K. Lee and H. Min, *Phys. Lett.* **B252** (1990) 79.
- [54] R. Jackiw and S-Y. Pi, *Phys. Rev. Lett.* **64** (1990) 2969 ; *Phys. Rev.* **D42** (1990) 3500.
- [55] S.K. Kim, W. Namgung, K.S. Soh and J.H. Yee, *Phys. Rev.* **D46** (1992) 1882.
- [56] S.K. Kim, K.S. Soh and J.H. Yee, *Phys. Rev.* **D42** (1990) 4139.
- [57] F. Wilczek and A. Zee, *Phys. Rev. Lett.* **51** (1983) 2250.
- [58] Y.-S. Wu and A. Zee, *Phys. Lett.* **B147** (1984) 325 ; A.M. Din and W. J. Zakrzewski, **B146** (1984) 341.
- [59] See for example, A. Lerda, *Anyons: Quantum Mechanics of Particles with Fractional Statistics*, Lecture Notes in Phys. m14 (Springer-Verlag, Berlin 1992) and references therein.

# 27. Chern Simons Field and Composite Bosons in the Quantum Hall system

R. Rajaraman \*

School of Physical Sciences Jawaharlal Nehru University  
New Delhi 110067, India

## Abstract

This is a brief pedagogical review of the use of the Chern Simons vector field and the associated singular gauge transformations in the theory of Quantum Hall Effect. We will primarily deal with the Fermion to Boson transformation to construct an order parameter field to characterise the Hall plateau states. The canonical exact Hamiltonian of this field is derived. The construction is generalised to include certain crucial non-unitary transformations which yield the well known and very successful Laughlin wavefunction as the Mean Field ground state of the condensate. The adaptation of the method to Jain's composite fermion theory is also briefly mentioned.

## 1 Introduction

The Quantum Hall System (by this we refer to quasi-two-dimensional layers of electrons trapped in the interface of semiconductors, at very high magnetic fields and very low temperatures) has revealed many remarkable features. The most well known of these is the occurrence of very special states at filling fractions  $\nu$  (defined as  $\frac{\hbar c \bar{\rho}}{e B}$ , where  $\bar{\rho}$  is the mean electron density and  $B$  the applied magnetic field) which are integers or certain odd denominator fractions. The special states at these filling fractions  $\nu_i$  lead to remarkably flat plateaus in Hall conductivity for a range of fillings around these values, with the conductivity (in units of  $\frac{e^2}{h}$ ) taking values exactly equal to  $\nu_i$  to an accuracy of 1 in  $10^7$ ! Associated with these plateaus in Hall conductivity is also a very steep drop in diagonal resistivity. These remarkable features are very universal in that they don't seem to depend on the details of the material, the extent (within some limits) and nature of the impurities etc. It was recognised that the electrons in these special Quantum Hall (QH) states form an incompressible fluid and Laughlin proposed a very simple set of trial wavefunctions to describe the ground states and lowlying excitations of these systems [1]. For some reviews of these phenomena and related basic theory, see Prange and Girvin [2], MacDonald, [3], Karlhede et al [4], and Stone [5].

That these special QH states correspond to some form of an ordering was first elucidated by Girvin and MacDonald [6] who showed that the Laughlin wavefunctions exhibit a non-trivial form of off-diagonal long range order. More precisely, they showed that there exists a composite operator of the fundamental Fermi fields that obeys Bose statistics and whose off-diagonal density matrix is algebraically long ranged in the Laughlin states [7]. This was an extremely important observation for it opened up the possibility of a Landau-Ginzburg description of the QHE in terms of an order parameter field, thereby bringing it within reach of more systematic computations.

In the phenomenon of superconductivity, the bosonic order parameter field is constructed from a product of two electron operators, reflecting the physics that the phenomenon arises from the condensation of Cooper pairs. No such pairing seems to be indicated in the physics of the QH system. However in two space dimensions, there is another way of constructing Bosons out of Fermions, and that is by using the "anyon" or Chern Simons (CS) transformation, for the special

---

\*Email: doug@jnu.ernet.in

case when the anyon angle is an odd multiple of  $\pi$  [8] [9]. Physically this amounts to making a composite of the electron with an odd number of flux tubes which then behaves like a Boson under exchange. It turns out that this particular way of constructing a Boson does correctly yield the field that seems to be getting ordered in the plateaus of the QH system, as anticipated by Girvin and MacDonald [6].

The bulk of this review will be devoted to the quantum field theoretic procedure for constructing such Quantum Hall order parameter fields and describing their dynamics. Before we do that however, it will be useful to recapitulate the Anyon-Chern Simons construction at the first quantised level, which we describe in the next section.

## 2 CS Interaction at the First Quantised Level

The mechanism of altering the effective Statistics of particles in two dimensions by multiplying them with appropriate phase factors, or equivalently making them interact with a statistical CS field has been well known for about two decades (for reviews see Wilczek [10]) and Sumathi Rao [11]). Therefore we will not dwell on the details of the basic ideas and move on to its application to the QH problem, following the work of Zhang, Hansson and Kivelson [12].

Consider a wavefunction of two electrons in two dimensions, antisymmetric under the exchange of the particles :

$$\psi(\vec{r}_1, \vec{r}_2) = -\psi(\vec{r}_2, \vec{r}_1) \quad (1)$$

Define a new function  $\phi$  by

$$\phi(\vec{r}_1, \vec{r}_2) \equiv e^{-im\theta_{12}} \psi(\vec{r}_1, \vec{r}_2) \quad (2)$$

where  $\theta_{12}$  is the angle (on the two dimensional plane) of the vector  $\vec{r}_{12} = \vec{r}_1 - \vec{r}_2$ . Since under the exchange  $1 \leftrightarrow 2$ ,  $\theta_{12} \rightarrow (\theta_{12} + \pi)$ , the field  $\phi$  picks up a factor  $e^{-im\pi}$  this amounts, for an arbitrary value of  $m$  to fractional or Anyon statistics. When  $m$  is an odd-integer,  $\phi$  corresponds to a pair of Bosons, while for even-integer values of  $m$  it continues to correspond to Fermions. For the present let us take  $m$  to be an odd-integer, so that  $\phi(\vec{r}_1, \vec{r}_2)$  is a bosonic wavefunction.

One can clearly generalise this to  $N$  particles. If  $\theta_{ij}$  is the angle of the vector  $\vec{r}_{ij} = \vec{r}_i - \vec{r}_j$  and  $\psi(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N)$  represents the wavefunction of  $N$  fermions, then,

$$\phi(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N) \equiv \left( \exp\left(-im \sum_{i < j} \theta_{ij}\right) \right) \psi(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N) \quad (3)$$

will, for odd  $m$ , be symmetrical under the exchange of any pair  $(ij)$ .

This phase factor will clearly induce an extra interaction in the Hamiltonian governing the function  $\phi$ , as compared to the one governing  $\psi$ . Let the fermionic wavefunction  $\psi$  be governed by a Hamiltonian

$$H_F = \sum_i^N -\frac{\hbar^2}{2\mu} \vec{D}_i^2 + eA_0(x_i) + \sum_{i < j} V(r_{ij}) \quad (4)$$

where  $\mu$  is the mass,  $A_0$  is some external electrostatic field, say, due to impurities,  $V(r_{ij})$  is the interparticle interaction and  $\vec{D} = \vec{\nabla} - \frac{ie}{\hbar c} \vec{A}$ . Then the bosonic function  $\phi$  will be governed by

$$H_B = \sum_i -\frac{\hbar^2}{2\mu} \left( \vec{D}_i - \frac{ie}{\hbar c} \vec{a}(\vec{r}_i) \right)^2 + eA_0(x_i) + \sum_{i < j} V(r_{ij}) \quad (5)$$

where the statistical gauge field  $\vec{a}$  is given by the gradient of the phase :

$$\vec{a}(\vec{r}_i) = \frac{\hbar c}{e} m \sum_{i \neq j} \vec{\nabla}_i \theta_{ij} \quad (6)$$

One can check that the vector field  $\vec{a}$  is transverse

$$\vec{\nabla} \cdot \vec{a} = 0 \quad (7)$$

and obeys

$$\text{Curl } \vec{a}(\vec{r}) = -\epsilon^{ij} \partial_i a_j = -m(\hbar c/e) \rho(\vec{r}) \quad (8)$$

where  $\rho(\vec{r})$ , the density, is just

$$\rho(\vec{r}) \equiv \sum_i \delta(\vec{r} - \vec{r}_i) \quad (9)$$

Note that even though (3) looks like an innocuous gauge transformation, it is singular. The associated "gauge" field, the Chern Simons vector field defined in (6) will yield non zero magnetic field as evident from eq (8).

In short, one can construct a bosonic (symmetric) wavefunction from a starting fermionic one at the cost of adding an interaction with a statistical gauge field  $\vec{a}$ . Note that with the constraints (7 and 8)  $\vec{a}$  is not an independent field. It is fully determined by the sources at  $\vec{r}_i$ . This vector field  $\vec{a}$  has come to be called the Chern Simons field because the constraint (8) can be encoded into the Lagrangian formulation by using terms in the Action of involving the Chern Simons 3-form:

$$S_{CS} = \int d^3x \left[ \kappa \epsilon^{\mu\sigma\lambda} a_\mu \partial_\sigma a_\lambda - e a_0 \rho \right] \quad (10)$$

where  $\kappa = \frac{e^2}{2m\hbar c}$  and  $\mu, \sigma$  and  $\lambda$  vary between 0, 1 and 2.

Notice from the definition of the filling fraction  $\nu$  that the external magnetic field obeys

$$B = \text{curl } \vec{A} = \frac{\hbar c \bar{\rho}}{e\nu} \quad (11)$$

which may be compared to eq(8) for  $\text{Curl } \vec{a}$ . When the filling fraction happens to be  $\nu = 1/m$  where  $m$  is the odd integer used in the CS transformation (3), we can see that the CS field constraint (8) in the mean-field approximation  $\rho(\vec{r}) = \bar{\rho}$  becomes

$$\text{curl } \vec{a} = \frac{m\hbar c \bar{\rho}}{e} = -\text{curl } \vec{A} \quad (12)$$

while

$$\vec{\nabla} \cdot \vec{a} = 0 = -\vec{\nabla} \cdot \vec{A} \quad (13)$$

Hence we can set  $\vec{a} = -\vec{A}$  in the Mean Field (MF) approximation. This in turn means that in the bosonic Hamiltonian (5), the covariant derivative becomes the ordinary gradient in the MF limit. Then in the absence of external electrostatic fields and inter-particle interactions taken to be cancelled in the MF limit by a neutralising background, (i.e. take  $A_0 = 0 = V(r_{ij})$ ), the bosonic Hamiltonian has a ground state solution

$$\phi(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N) = 1 \quad (14)$$

which can be interpreted as a translationally invariant condensate of zero-momentum particles.

Thus by using the phase transformation 3 and associated interaction with the CS field  $\vec{a}$  one has not only gotten a bosonic wavefunction, but also ground state where this wavefunction is a constant in space (a condensate of the bosons)! The corresponding wavefunction for the electrons, in this condensate ground state, obtained by inverting the CS phase transformation (3) will be

$$\psi(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N) = \exp\left(im \sum_{i < j} \theta_{ij}\right) \quad (15)$$

At this stage we recall that, long before the above developments, we had the famous Laughlin wavefunction [1] which describes to great accuracy the QH ground state at filling fractions  $\nu = 1/m$  with  $m$  odd. This wavefunction is :

$$\psi_L(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N) = \prod_{i < j} (z_i - z_j)^m e^{-\sum_i \frac{|z_i|^2}{4l^2}} \quad (16)$$



where  $z_i = x_i + iy_i$  is the complex coordinate associated with the point  $\vec{r}_i$ .  $l = \sqrt{\frac{\hbar c}{eB}}$  is the magnetic length. If we compare our mean field solution (15) with this Laughlin wavefunction we see that the former is precisely the phase of the latter.

That the CS method produces also the phase of the Laughlin wavefunction is another indication the the electron-CS-flux composite (the boson above) represents a good bit of the correct physics of the QH system at fractions  $\nu = 1/m$ .

However, the modulus of the Laughlin wavefunction  $\psi_L$  contains its all- important zeroes  $|(z_i - z_j)|^m$  when any two electrons approach each other, crucial for reducing the Coulomb energy between them. And the gaussian factors, also part of the modulus of  $\psi_L$ , are vital for its normalisibility. Of course Zhang [13] did show that inclusion of fluctuations around the mean field solutions in the Random Field Approximation ends up reproducing the full Laughlin wavefunction including its modulus. But that means that fluctuations are not producing small corrections to the MF result. They are doing considerable violence to the MF wavefunction by scooping out holes in it when two particles approach one another, apart from introducing strong damping (gaussian) factors when any coordinate  $z_i$  becomes large. It would be nice if one could generalise the Fermion-Boson transformation in some way so that the the entire laughlin wave function (16) (and not just its phase) comes out already in the Mean Field limit.

The kernel of such a possibility lies in noting that eq(3) is not the only unique way to generate symmetric wavefunctions from anti symmetric ones. We can multiply the right hand side by any symmetric function of the coordinates and  $\phi$  will still be symmetric. One could see if the choice of a suitable symmetric pre-factor can lead to the full Laughlin wavefunction corresponding to the MF condensate. Such a procedure was set up at the quantum field theoretical level by the present author and Sondhi [14]. We will discuss that in the next section.

### 3 The Fermion-Boson Transformation in QFT

Let  $\Psi(\vec{x})$  be the full electron field operator obeying the equal-time anticommutation relations,

$$\begin{aligned} \{\Psi(\vec{x}), \Psi(\vec{x}')\} &= \{\Psi^\dagger(\vec{x}), \Psi^\dagger(\vec{x}')\} = 0 \\ \{\Psi(\vec{x}), \Psi^\dagger(\vec{x}')\} &= \delta^{(2)}(\vec{x} - \vec{x}') . \end{aligned} \quad (17)$$

The second quantized field theoretic Hamiltonian that describes our system is,

$$H = \int d^2x \left[ \Psi^\dagger(\vec{x}) \left( \frac{-\hbar^2}{2\mu} \vec{D}^2 + eA_0 \right) \Psi(\vec{x}) \right] + \frac{1}{2} \int \int d^2x d^2x' \delta\rho(\vec{x}) V(\vec{x} - \vec{x}') \delta\rho(\vec{x}') . \quad (18)$$

Recall that here  $\vec{D} \equiv \vec{\nabla} - i\frac{e}{\hbar c} \vec{A}$  with  $\vec{A}$  being the vector potential of the uniform field ( $B = \nabla \times \vec{A}(\vec{x})$ ) and  $\rho(\vec{x}) \equiv \Psi^\dagger(\vec{x})\Psi(\vec{x})$  is the density operator whose deviation from its mean value  $\bar{\rho}$  is  $\delta\rho(\vec{x})$ .

In order to recast this as a bosonic problem we need a set of canonical bosonic operators which we construct in the next sub-section.

#### 3.1 Bosonic Operators

Consider the pair of operators  $\Phi(\vec{x})$  and  $\Pi(\vec{x})$ , defined by

$$\begin{aligned} \Phi(\vec{x}) &\equiv e^{-J(\vec{x})} \Psi(\vec{x}) \\ \Pi(\vec{x}) &\equiv \Psi^\dagger(\vec{x}) e^{J(\vec{x})}, \end{aligned} \quad (19)$$

where,

$$J(\vec{x}) \equiv m \int d^2x' [\rho(\vec{x}') \log(z - z')] - \frac{|z|^2}{4l^2}, \quad (20)$$

and  $m$  is an odd integer.

It is clear that in the field theoretic version of the purely phase transformation (3) used by ZHK [12],  $J$  would be chosen to contain only the phase of  $(z - z')$ , i.e.  $\text{Im} \log(z - z')$  in place of the  $\log(z - z')$  in our definition (20) and with no  $|z|^2$  term. Consequently their  $\Pi = \Phi^\dagger$ . The field theoretic description of the ZHK work can be read off from the equations below by making these replacements. We will proceed using the generalised definition (20)

The Bosonic operator  $\Phi$  in (19) was in fact introduced first by Read [15] long before ref[14] as the appropriate order parameter field for the Qu Hall ground states. He also proceeded to write the classical field equation obeyed by the expectation value of this Bose field under the Lowest Landau Level and other approximations. Here we adopt Read's operator, but describe a full-fledged quantum field theory for it at the operator level without any approximations in the first instance. As a first step we have defined the field  $\Pi$  in (19) as the canonical conjugate of  $\Phi$ . Eventually we will write down the quantum field (18) entirely in terms of Bose fields by changing variables from  $\Psi$  and  $\Psi^\dagger$  to  $\Phi$  and  $\Pi$ .

Evidently,  $\Phi$  and  $\Pi$  are not hermitian conjugates as  $J$  has both hermitian and anti-hermitian pieces; in fact,

$$\Pi(\vec{x}) = \Phi^\dagger(\vec{x}) e^{J(\vec{x}) + J^\dagger(\vec{x})}. \quad (21)$$

Nevertheless, as we now show, they are canonically conjugate Bose fields.

To this end note that the only operator appearing in  $J(\vec{x})$  is the electron density  $\rho(\vec{x}) = \Psi^\dagger(\vec{x})\Psi(\vec{x})$ , which obeys the commutation relation,

$$[\rho(\vec{x}), \Psi(\vec{x}')] = -\Psi(\vec{x}) \delta^2(\vec{x} - \vec{x}'). \quad (22)$$

Using this one can obtain the following identities:

$$\begin{aligned} e^{-J(\vec{x})} \Psi(\vec{x}') &= (z - z')^m \Psi(\vec{x}') e^{-J(\vec{x})} \\ \Psi^\dagger(\vec{x}') e^{-J(\vec{x})} &= (z - z')^m e^{-J(\vec{x})} \Psi^\dagger(\vec{x}'). \end{aligned} \quad (23)$$

It is then straightforward to verify using (23) that

$$[\Phi(\vec{x}), \Phi(\vec{x}')] = [\Pi(\vec{x}), \Pi(\vec{x}')] = 0 \quad (24)$$

while

$$[\Phi(\vec{x}), \Pi(\vec{x}')] = \delta^2(\vec{x} - \vec{x}'). \quad (25)$$

Thus, despite the presence of non-unitary factors in their definition in Eq. (19), the fields  $\Phi$  and  $\Pi$  form a pair of mutually canonical Bose fields. However, in contrast to standard charged scalar field theories, here  $\Pi$  is not equal to  $\Phi^\dagger$ , instead they obey the more complicated relation in Eq. (21). This fact, a consequence of the non-unitary transformation in (19), has to be borne in mind in doing manipulations with our theory.

Nevertheless, notice that the fermion density  $\rho$ , when written in terms of  $\Pi$  and  $\Phi$  still has the standard bosonic form

$$\rho(\vec{x}) = \Psi^\dagger(\vec{x})\Psi(\vec{x}) = \Pi(\vec{x})\Phi(\vec{x}). \quad (26)$$

Thus, if  $N \equiv \int d^2x \rho$  is the number operator, then

$$[N, \Pi(\vec{x})] = \Pi(\vec{x}), \quad (27)$$

i.e. the operator  $\Pi(\vec{x})$  creates one extra composite boson, and the number of composite bosons is the same as the number of the original fermions.

### 3.2 The Hamiltonian

Consider the action of the covariant derivative on the electron field. We have,

$$\begin{aligned} \vec{D}\Psi(x) &= \vec{D}(e^{J(\vec{x})}\Phi(\vec{x})) \\ &= \left(\vec{\nabla} - i\frac{e}{\hbar c}\vec{A}(\vec{x})\right) (e^{J(\vec{x})}\Phi(\vec{x})) \end{aligned}$$

$$\begin{aligned}
&= e^{J(\vec{x})} \left( \vec{\nabla} - i \frac{e}{\hbar c} \vec{A}(\vec{x}) + \vec{\nabla} J(\vec{x}) \right) \Phi(\vec{x}) \\
&= e^{J(\vec{x})} \left( \vec{D} - i \frac{e}{\hbar c} \vec{v}(\vec{x}) \right) \Phi(\vec{x})
\end{aligned} \tag{28}$$

where,

$$\vec{v}(\vec{x}) \equiv i \frac{\hbar c}{e} \vec{\nabla} J(\vec{x}) . \tag{29}$$

Hence,

$$D^2 \Psi = e^J \left( \vec{D} - i \frac{e}{\hbar c} \vec{v} \right)^2 \Phi \tag{30}$$

Inserting this into the starting Hamiltonian (18), and using Eqs. (19) and (26) we get,

$$\begin{aligned}
H &= \int d^2 x \left[ \Pi(\vec{x}) \left( \frac{-\hbar^2}{2\mu} \left( \vec{\nabla} - i \frac{e}{\hbar c} (\vec{A} + \vec{v}) \right)^2 + e A_0 \right) \Phi(\vec{x}) \right] \\
&+ \frac{1}{2} \int \int d^2 x d^2 x' \delta \rho(\vec{x}) V(\vec{x} - \vec{x}') \delta \rho(\vec{x}')
\end{aligned} \tag{31}$$

This Hamiltonian, the auxilliary definitions (26) and (29) and the commutators (24, 25) together define a purely bosonic problem that is fully equivalent to our original fermion problem .

The vector field  $\vec{v}$  appearing in 31 above is constrained in terms of the density by Eq. (29), where  $J(\vec{x})$  is defined in (20). Since this  $J(\vec{x})$  involves more than just the phase of  $(z - z')$ , this field  $\vec{v}$  is not the field theoretic version of statistical Chern-Simon gauge field  $\vec{a}$  used in the last section. Because  $J(\vec{x})$  has real parts,  $\vec{v}$  is a complex vector field. However, we will see now that  $\vec{v}$  is simply related to  $\vec{a}$ .

In field theoretic formulation the Chern-Simons field is defined as

$$\vec{a}(\vec{x}) = \frac{-m\hbar c}{e} \vec{\nabla}_x \int d^2 x' \rho(\vec{x}') \text{Im} \log(z - z') , \tag{32}$$

or equivalently

$$b \equiv \nabla \times \vec{a} = -m\phi_0 \rho \tag{33}$$

where  $\phi_0 \equiv \frac{\hbar c}{e}$  is the flux quantum. Now, the function  $\log z$  obeys the Cauchy-Riemann conditions away from  $z = 0$ , which can be written as

$$\vec{\nabla}(\text{Re} \log z) = \vec{\nabla}(\text{Im} \log z) \times \hat{k} \tag{34}$$

where  $\hat{k}$  is a unit vector perpendicular to the plane. Using this we get,

$$\begin{aligned}
\vec{v}(\vec{x}) &= \frac{i\hbar c}{e} \vec{\nabla} J(\vec{x}) \\
&= \frac{i\hbar c}{e} \vec{\nabla}_x m \int d^2 x' [\rho(\vec{x}') (\text{Re} \log(z - z') + i \text{Im} \log(z - z'))] - \frac{|z|^2}{4l^2} \\
&= \vec{a}(\vec{x}) + i \hat{k} \times \vec{a}_{cs}(\vec{x}) - \frac{i\hbar c}{e} \frac{\vec{x}}{2l^2} .
\end{aligned} \tag{35}$$

Note that the last term in the above equation is just a c-number term involving the coordinate vector  $\vec{x}$ . The density dependent operator part of  $\vec{v}$  is present *entirely* through  $\vec{a}$ .

## 4 The Chern Simons Action and its Field Equations

The constraint (33) relating  $\vec{a}$  to the density  $\rho$  and the transversality condition can be implemented by the usual device of introducing a Lagrange multiplier fields in the action formalism. This action, as mentioned already, involves the Chern Simons 3-form  $ada$  (in form notation). In detail, the Action is

$$S = \int d^2 x dt \left[ \Pi (i\hbar \partial_t - e a_0) \Phi - \frac{e}{2m\phi_0} \epsilon^{\mu\nu\sigma} a_\mu \partial_\nu a_\sigma + \lambda \vec{\nabla} \cdot \vec{a} \right] - \int dt H \tag{36}$$

where  $H$  is the bosonized Hamiltonian in (31), and  $a^\mu$  ( $\mu = 0, 1, 2$ ) is the 3-vector  $(a_0, \vec{a})$ .

There is, however, a subtlety in this procedure which does not arise in ZHK's construction and has to do with the gauge invariance of the resulting action and hence the freedom to pick gauges different from transverse gauge. First, note that the action is manifestly invariant with respect to gauge changes of the external field, i.e. the transformations,

$$\begin{aligned}\vec{A} &\rightarrow \vec{A} - \frac{\hbar c}{e} \vec{\nabla} \Lambda(\vec{x}, t) \\ A_0 &\rightarrow A_0 + \frac{\hbar}{e} \partial_t \Lambda(\vec{x}, t) \\ \Phi &\rightarrow e^{-i\Lambda(\vec{x}, t)} \Phi \\ \Pi &\rightarrow e^{+i\Lambda(\vec{x}, t)} \Pi.\end{aligned}\quad (37)$$

However, the invariance with respect to gauge changes of the Chern-Simons field is more restricted. Gauge transformations of the form (37) with  $(A_0, \vec{A})$  replaced by  $(a_0, \vec{a})$  leave the action invariant only if  $\Lambda(\vec{x}, t)$  is independent of  $\vec{x}$ , i.e. if they do not involve the spatial gauge field at all. In addition, there is a class of modified gauge transformations for the spatial components that do not involve the temporal component of the gauge field and have the following form. Let  $f(z) = u(\vec{x}) + iw(\vec{x})$  be an analytic function of  $z$ . Then the action is invariant under,

$$\begin{aligned}\vec{a} &\rightarrow \vec{a} - \frac{\hbar c}{e} \vec{\nabla} w(\vec{x}) \\ \Phi &\rightarrow e^{-f(z)} \Phi \\ \Pi &\rightarrow e^{+f(z)} \Pi.\end{aligned}\quad (38)$$

(The variation of the gauge field implies that  $J \rightarrow J + f$  which ensures that the constraint Eq. (21) is preserved.) The implications of this feature of our theory, in particular the significance of the modified gauge invariance (38) and its possible connection to work on  $W_\infty$  algebras [16], remain a subject for future work.

The field equations arising from the Action (36) are a "non-linear Schrödinger equation" [17],

$$\begin{aligned}(i\hbar\partial_t - e(a_0 + A_0))\Phi(\vec{x}) &= -\frac{\hbar^2}{2\mu} \left[ \vec{\nabla} - \frac{ie}{\hbar c} \left( \vec{A} + \vec{a} + i\hat{k} \times \vec{a} - \frac{i\hbar c}{e} \frac{\vec{x}}{2l^2} \right) \right]^2 \Phi(\vec{x}) \\ &+ \left( \int d^2x' V(\vec{x} - \vec{x}') \delta\rho(\vec{x}') \right) \Phi(\vec{x}),\end{aligned}\quad (39)$$

along with the modified Chern-Simons field-current identities,

$$\vec{\nabla} \times \vec{a} = -m\phi_0 \Pi \Phi \quad (40)$$

$$\vec{\nabla} \cdot \vec{a} = 0 \quad (41)$$

$$\hat{k} \times (-\partial_0 \vec{a} - \vec{\nabla} a_0) = \frac{m\phi_0}{c} (\vec{j} - i\hat{k} \times \vec{j} + \nabla\lambda) \quad (42)$$

where,

$$\vec{j} = \frac{\hbar}{2\mu i} [\Pi (\vec{\mathcal{D}} \Phi) - (\vec{\mathcal{D}} \Pi) \Phi] \quad (43)$$

$$\vec{\mathcal{D}} \equiv \vec{\nabla} - \frac{ie}{\hbar c} \left( \vec{A} + \vec{a} + i\hat{k} \times \vec{a} - \frac{i\hbar c}{e} \frac{\vec{x}}{2l^2} \right) \quad (44)$$

Although this current  $\vec{j}$  does not look manifestly hermitian, it is in fact just the usual hermitian electron current operator, as can be verified by rewriting it in terms of the Fermi fields.

#### 4.1 Exact classical "Ground State" Solution

These field equations have a simple exact solution for the situation where the external electric potential  $A_0$  is absent and the uniform magnetic field  $B = \nabla \times \vec{A}$  is chosen so that the filling fraction is

$$\nu \equiv \frac{\bar{\rho}\phi_0}{B} = \frac{1}{m}, \quad (45)$$

where  $m$  is the odd integer in the fermion to boson transformation function  $J$  defined in (20). The solution describes a homogeneous state and is given by the fields,

$$\begin{aligned} \Phi(\vec{x}) &= \Pi(\vec{x}) = \sqrt{\bar{\rho}} \\ \vec{a}_{\bar{\rho}}(\vec{x}) &= \frac{m\bar{\rho}\phi_0}{2}(\vec{x} \times \hat{k}) \\ a_0 &= 0 = \lambda. \end{aligned} \quad (46)$$

In order to verify that this is indeed a solution, we begin by noting that density  $\rho = \Pi\Phi$  equals its mean value  $\bar{\rho}$  everywhere. Hence  $\delta\rho$  is zero, removing the last term in (39). It follows that a constant  $\Phi$  solves (39) provided the gauge fields that enters the covariant derivatives vanish. For the temporal gauge field this is trivially true. For the spatial gauge field we note that,

$$\begin{aligned} \vec{a}_{\bar{\rho}}(\vec{x}) &= -\frac{1}{2}m\phi_0\bar{\rho}\hat{k} \times \vec{x} \\ &= -\frac{1}{2}B\hat{k} \times \vec{x} \\ &= -\vec{A}, \end{aligned} \quad (47)$$

and hence

$$\vec{a} + \vec{A} = 0 \quad (48)$$

This condition for picking out uniform states, that the Chern-Simons field at mean density cancels the external field  $\vec{A}$ , is already known from [12]. But the statistical gauge field appearing in the covariant derivative in our Lagrangian and field equation (36 and 39) is not just  $\vec{A} + \vec{a}$ . It also contains imaginary pieces. However, we also have the additional result that

$$\begin{aligned} \hat{k} \times \vec{a}_{\bar{\rho}} &= \frac{B}{2}\vec{x} \\ &= \frac{c\hbar}{e} \frac{\vec{x}}{2l^2} \end{aligned} \quad (49)$$

This last equality tells us that the extra imaginary pieces of the statistical gauge field, i.e. the third and fourth terms in Eq. (35), also cancel one another. Altogether, we have, for  $\rho = \bar{\rho}$ ,

$$\vec{A} + \vec{a}_{\bar{\rho}} + i\hat{k} \times \vec{a}_{\bar{\rho}} - i\frac{c\hbar}{e} \frac{\vec{x}}{2l^2} = 0 \quad (50)$$

and hence the forms (46) satisfy Eq. (39).

It is also straightforward to verify that the forms (46) solve the field-current identity (3.5). Finally, readers concerned about the consistency of the solutions for  $\Phi$  and  $\Pi$  should note that for our solutions  $J + J^\dagger = 0$ .

### 5 The Mean Field Wavefunction

The classical solution (46) obtained in the previous section contains the boson field  $\Phi$  which is space-time independent, has a uniform phase and a non-vanishing amplitude everywhere. This solution can be viewed as the ground state expectation value of an ideal condensate of the composite bosons

We now show that N-particle projection of this condensate ground state is, in the first quantized fermionic representation, exactly the Laughlin state. To see this note from Eq. (27) that in our bosonized formulation, an N-particle state is obtained by the action of N powers of  $\Pi$  on the vacuum. Hence the translationally invariant ground state, where all the bosons have condensed into the  $k = 0$  mode, has the (arbitrarily normalized) form,

$$|N\rangle_{MF} = \frac{1}{N!} \left[ \int d^2x \Pi(\vec{x}) \right]^N |O\rangle \quad (51)$$

where  $|O\rangle$  is the no particle (vacuum) state, and  $N, V \rightarrow \infty$  with  $\frac{N}{V} = \bar{\rho}$ . The first quantized *electron* wavefunction associated with this state is

$$\begin{aligned} \psi_{MF}(\vec{y}_1, \vec{y}_2, \dots, \vec{y}_N) &= \langle O | \Psi(\vec{y}_1) \dots \Psi(\vec{y}_N) | N \rangle_{MF} \\ &= \frac{1}{N!} \langle O | e^{J(\vec{y}_1)} \Phi(\vec{y}_1) e^{J(\vec{y}_2)} \Phi(\vec{y}_2) \dots e^{J(\vec{y}_N)} \Phi(\vec{y}_N) \\ &\quad \times \int d^2x_1 \Pi(\vec{x}_1) \int d^2x_2 \Pi(\vec{x}_2) \dots \int d^2x_N \Pi(\vec{x}_N) | O \rangle . \end{aligned} \quad (52)$$

Now we can use the identity,

$$\Phi(\vec{y}_1) e^{J(\vec{y}_2)} = e^{J(\vec{y}_2)} \Phi(\vec{y}_1) (z_1 - z_2)^m, \quad (53)$$

to move all the factors of  $e^J$  to the left, which yields

$$\begin{aligned} \psi_{MF}(\vec{y}_1, \vec{y}_2, \dots, \vec{y}_N) &= \frac{1}{N!} \prod_{i < j} (z_i - z_j)^m \langle O | e^{\sum_i J(\vec{y}_i)} \Phi(\vec{y}_1) \dots \Phi(\vec{y}_N) \\ &\quad \times \int d^2x_1 \Pi(\vec{x}_1) \int d^2x_2 \Pi(\vec{x}_2) \dots \int d^2x_N \Pi(\vec{x}_N) | O \rangle . \end{aligned} \quad (54)$$

Next we use Wick's theorem for the product of the  $\Phi$ 's and  $\Pi$ 's and note that the former annihilate the vacuum to obtain

$$\psi_{MF}(\vec{y}_1, \vec{y}_2, \dots, \vec{y}_N) = \prod_{i < j} (z_i - z_j)^m \langle O | e^{\sum_i J(\vec{y}_i)} | O \rangle . \quad (55)$$

Finally, as the vacuum has no particles, only the gaussian factor in  $J$  contributes and we have the result,

$$\psi_{MF}(\vec{y}_1, \vec{y}_2, \dots, \vec{y}_N) = \prod_{i < j} (z_i - z_j)^m e^{-\sum_i \frac{|z_i|^2}{4l^2}} . \quad (56)$$

Thus the mean field state directly yields the complete Laughlin wavefunction. This is in contrast to the bosonized theory of ZHK, where the mean field state contains only the correct phase of the Laughlin wavefunction and not its zeroes or the gaussian factor, fluctuations are included.

## 6 Further Developments

The main theme of this review, as indicated in the title, was the Chern Simons method for constructing bosons which are then shown to condense in the ground state at fillings  $\nu = 1/m$  with odd  $m$ . We will conclude by briefly mentioning some closely related further developments, with references where fuller details may be found.

## 6.1 Composite Fermions

The bosons discussed above can be viewed as a composite of the original electron and an odd number ( $m$ ) of fluxons. These are sometimes called composite bosons. It is evident from the way the composite operators were constructed in eq (19) and their commutation property studied in eq (24) that if the integer  $m$  in (20) were even instead of odd, the composite operators would obey canonical anti-commutation relations instead of commutation relations. We would thus have composite fermions instead of composite bosons when  $m$  is even.

While such composite fermions would clearly be of no relevance as candidates for an order parameter, they turn out to be very useful in a different way. It may be recalled that about 10 years ago Jain [18] offered an understanding of fractional Qu Hall plateaus based on certain composite fermions made of even number of fluxons and electrons. His theory addressed Hall plateaus at electron fillings  $\nu = \frac{p}{(1+2pk)}$  where  $p$  and  $k$  are integers. He related these states to Qu Hall states of his composite fermions at integer fillings  $p$ . He also argued that the wavefunction of the fractional filling electron states  $\psi_\nu(z_1, z_2, \dots, z_N)$  can be related to the composite fermion wavefunction  $\phi_p(z_1, z_2, \dots, z_N)$  by the equation

$$\psi_\nu(z_1, z_2, \dots, z_N) = \prod_{i < j} (z_i - z_j)^{2k} e^{-\sum_i (2k\nu) \frac{|z_i|^2}{4}} \cdot \phi_p(z_1, z_2, \dots, z_N) \quad (57)$$

It was shown by Lopez and Fradkin [19] that these ideas, which Jain introduced through powerful intuition and physical arguments, can be also be systematically derived using the Chern Simons transformation we have discussed here, with an even value of the integer  $m = 2k$ . Once again Lopez and Fradkin used only the pure phase transformation involving the relative angle, akin to the ZHK work, and obtained, at the mean field level, only the phase of the Jain prefactors. However, a transformation of the kind described here, using the full exponent  $J(x)$  as in eq(20), was done by the present author [20] which yields all of Jain's prefactors in the relation (57). (See also Wu and Yu [21]).

A generalisation of the composite fermion picture for Quantum hall systems in double layers was also done by Lopez and Fradkin [22] and also by the author [20].

## 6.2 Filling Factor 1/2

The case  $\nu = 1/2$  does not correspond to one of the Hall plateaus. But at this filling many other interesting things happen. This case has been studied extensively and by different competing groups using different approaches. See for instance Murthy and Shankar [23], Halperin *et al* [24], and references therein. We can examine this system using our Chern Simons transformation. Suppose the transformation 19 were employed with the integer  $m = 2$  at filling  $\nu = 1/2$ . Then, on the one hand the new composite field  $\Phi$  would be a fermi field, and at the same time in the mean field approximation the external electromagnetic potential would be cancelled by the Chern Simons field exactly as in eq(48). The composite fermions would then be free fermions, and could be described by a Fermi sea Slater determinant of plane wave states. The wavefunction of the system in electronic coordinates can then be written in this Mean Field approximation by multiplying the Fermi Sea wavefunction  $|FS\rangle$  by the prefactor for  $m = 2$  to give

$$\psi_{1/2} = \prod_{i < j} (z_i - z_j)^2 e^{-\sum_i \frac{|z_i|^2}{4}} \cdot |FS\rangle \quad (58)$$

Such a wavefunction was first proposed by Rezayi and Read [25].

## 6.3 Fluctuations

Our main theme in this article was the setting up of a quantum field theory for a suitably defined order parameter field for the quantum Hall effect. The field theory admits mean-field states at the fractions  $\nu = 1/m$  that are ideal condensates in the Bose language and correspond exactly to

the Laughlin states in terms of the electrons. In order to treat fluctuations about these states and to calculate the spectra of the various collective modes it is necessary to perturb about the mean field Hamiltonian. In our formulation, the mean-field Hamiltonian has the simple form,

$$H_{MF} = -\frac{\hbar^2}{2\mu} \int d^2x \Pi(\vec{x}) \nabla^2 \Phi(\vec{x}) \quad (59)$$

and hence its eigenstates are all known exactly. Nevertheless,  $H_{MF}$  is non-hermitian and hence states with different energies are not necessarily orthogonal. (The full Hamiltonian is perfectly hermitian; however the mean-field theory dictates that we decompose it as the sum of two non-hermitian pieces.) This requires then, that the perturbation theory explicitly take account of the non-orthogonality and that we possess tractable expressions for the overlaps between different states. For some work in this direction see Wu and Yu [21].

It is a pleasure to thank Dr S.L.Sondhi for collaboration and numerous illuminating discussions on this subject.

## References

- [1] R. B. Laughlin, Phys. Rev. Lett. **50** 1395 (1983).
- [2] "Quantum Hall Effect" edited by R.E.Prange and S.M. Girvin, Springer, (New York), (1990).
- [3] "The Quantum Hall Effect: A Perspective", edited by A.H.MacDonald, Kluwer, (Boston), (1989).
- [4] A.Karlhede, S.A.Kivelson and S.L.Sondhi, "The Quantum Hall Effect – The Article" Lectures at the 9th Jerusalem Winter School on Theoretical Physics, (1992)
- [5] "Quantum Hall Effect", (Ed M. Stone), World Scientific, Singapore, (1992).
- [6] S. M. Girvin and A. H. MacDonald, Phys. Rev. Lett. **58**, 1252 (1987).
- [7] Or rather, there is a set of operators, each element of which is long ranged in a different state. For linguistic simplicity we generally refer to all the elements of the set in the singular, e.g. "the bosonic operator".
- [8] J.M.Leinass and J.Myrheim, Il Nuovo Cimento, **37**, 1, (1977).
- [9] F. Wilczek, Phys. Rev. Lett. **49**, 957, (1982).
- [10] Frank Wilczek, "Fractional Statistics and Anyon Superconductivity", World Scientific, (Singapore), (1990).
- [11] Sumathi Rao, "Anyon Primer", in "Models and Techniques of Statistical Physics", M. Bhattacharya (Ed.), Narosa Publishers, (New Delhi), (1995); TIFR preprint TIFR/TH/92-18: hep-th-9209066.
- [12] S. C. Zhang, T. H. Hansson and S. Kivelson, Phys. Rev. Lett. **62**, 82, (1989).
- [13] S. C. Zhang, Int. J. Mod. Phys. B **6**, 25 (1992).
- [14] R.Rajaraman and S.L.Sondhi, Int.J.Mod.Phys.B **7**, vol.10, 793, (1996).
- [15] N. Read, Phys. Rev. Lett. **62**, 86 (1989).
- [16] S. Iso, D. Karabali and B. Sakita, Phys. Lett. B **296**, 143 (1992); A. Cappelli, C. A. Trugenberger and G. R. Zemba, Nucl. Phys. B **396**, 465 (1993).
- [17] There is also a related equation for  $\Pi$ .



- [18] J.K.Jain, Phys. Rev. Lett. **63**, 199 (1989); Phys. Rev. B **40**, 8079 (1989).
- [19] A. Lopez and E. Fradkin, Phys. Rev. Lett. **69**, 2126 (1992) ; A. Lopez and E. Fradkin, Phys. Rev. B **44**, 5246 (1991).
- [20] R.Rajaraman, Phys.Rev B **56**, 6788, (1997).
- [21] Y-S Wu and Y.Yu , "Field Theory of Vortex like Composite Fermions, preprint cond-matt. 9608061 (1996).
- [22] A. Lopez and E. Fradkin, Phys. Rev. B **51**, 4347 (1995).
- [23] G. Murthy and R.Shankar, Phys. Rev. Lett. **79**, 4437 (1997); see also preprint cond matt 9802244 (1998)
- [24] B. I. Halperin, P. A. Lee and N. Read, Phys. Rev. B **47**, 7312 (1993).
- [25] E.Rezayi and N Read, Phys. Rev. Lett. **72**, 900, (1994).

## Part F : Methods Of Strong Interactions In QFT

- 28. Hadrons From QCD - Achievements And Prospects by Olivier Pene
- 29. QCD Sum Rules In Hadronic And Nuclear Physics by L.S.Kisslinger
- 30. Light-Front Dynamics by V.A.Karmanov
- 31. 3D-4D Interlinkage Of B-S Amplitudes - Unified View Of  $Q\bar{Q}$  And  
 $QQQ$  Dynamics by A.N.Mitra
- 32. The Harmonic Oscillator In Quantum Theory - A Powerful Bridge In Physics  
by Marcos Moshinsky



# 28. Hadrons from QCD: achievements and prospects

O. Pène<sup>a \*</sup>

<sup>a</sup>Laboratoire de Physique Théorique <sup>†</sup>

Université de Paris XI, Bâtiment 211, 91405 Orsay Cedex, France

## Abstract

We overview very briefly some main attempts to deduce the hadron properties from QCD. After a short reminder of quark models we recall the role played by the use of symmetries, effective Lagrangians and Wilson expansion. We then turn to lattice Monte-Carlo simulations, claim that for simple matrix elements their limitations are only technical and quote some of their important recent predictions. We indicate how lattices may establish a link between non-perturbative and perturbative QCD. Finally we throw a prospective look towards the expected achievements in the coming years.

## 1 Introduction

Quantum Chromodynamics (QCD) is today unanimously considered as the theory of the strong interactions. From the tiny compact formula of the QCD Lagrangian,

$$\mathcal{L}_{QCD} = -\frac{1}{4}G_{\mu\nu}^a G_{\mu\nu}^a + i \sum_q \bar{\psi}_q \gamma^\mu (D_\mu)_{ij} \psi_q^j - \sum_q m_q \bar{\psi}_q \psi_q, \quad (1)$$

where

$$G_{\mu\nu}^a = \partial_\mu A_\nu^a - \partial_\nu A_\mu^a + g_s f_{abc} A_\mu^b A_\nu^c, \\ (D_{\mu\nu})_{ij} = \delta_{ij} - ig_s \sum_a \frac{\lambda_{ij}^a}{2} A_\mu^a$$

one can derive successful predictions for an enormous amount of physical observations.

But “there is many a slip twix the cup and the lip”. The full derivation of physics from QCD is often a hard task. Of course QCD was born when asymptotic freedom has been discovered and it was understood that one can use perturbation theory, Feynman graphs, to compute processes which involve large energies, large momentum transfers, in other words small spatial distances. Under this respect QCD is somewhat similar to Quantum Electrodynamics.

Still, complementary to asymptotic freedom, is the small energy behaviour of QCD, more mysterious and original: confinement. QCD has this extraordinary specificity that the states of the theory, the hadrons (baryons and mesons), are *not similar* to the fundamental fields of eq. (1) namely the quarks and gluons. Any concrete process in QCD starts and ends with hadrons, even though any theoretical work starts from quarks and gluons.

It is therefore mandatory to learn how to describe hadrons from quarks and gluons, to learn how to compute hadron properties: masses, matrix elements, etc. from eq. (1). And this is extremely difficult. The realm of QCD is currently divided into perturbative and non-perturbative domains. In the former, the main tool is Feynman graphs. The latter can be subdivided into first the many-body non perturbative processes, soft scattering, small  $x$  physics, the domain of Regge theory, and

---

\*E.mail:pene@qcd.th.u-psud.fr

<sup>†</sup>Unité Mixte de Recherche - UMR 862,

second, the hadronic physics. The major tool in the latter is Monte-Carlo lattice QCD numerical simulations. However, this division of QCD into subdomains is largely artificial and illegitimate. The interconnections between perturbative and non-perturbative physics are omnipresent. As already stated, hard processes start and end with hadrons. Deep inelastic scattering incorporates perturbative coefficients and a non-perturbative structure functions, etc.

*QCD is one domain.* We would like to explore this unity the other way round. Instead of starting from perturbative processes and wondering how one can reinsert hadrons, we would like to start from hadrons, overview the methods which have been used, from quark models, effective Lagrangians, symmetries, to lattice. We would like to discuss in what sense lattice is exactly QCD, and then consider how bridges are thrown from lattice QCD towards other fields.

## 2 Quark models

QCD originated from quark models [1]: a “lego” game describing the hadron spectrum in terms of quarks. Very quickly it was proposed to describe hadrons as bound states of quarks satisfying some kind of Schrödinger equation, and dynamical properties, such as transition matrix elements of hadrons, were computed from these wave functions [2]. It was later realised that the “partons” seen in deep inelastic scattering were indeed quarks, and finally [3] the discovery of the asymptotic freedom of the Lagrangian (1) started up real QCD.

But quark models continued, mainly because there existed no other method to compute hadronic matrix elements. The discovery of charmonium opened a field where quark models can be in part derived from QCD. Many attempts [4] towards relativistic quark models have existed with good phenomenological properties. But none can be rigorously derived from QCD in a systematic way. Quark models remain *models* even though they are still indispensable today since it remains impossible to study, with any other method, processes which involve orbitally or radially excited resonances. Quark models are also the only ones to provide a complete description of hadron spectrum in a semi-quantitative way.

## 3 Symmetries and effective Lagrangians

### 3.1 Flavor symmetries

In view of the difficulty to derive hadron properties from QCD, the first move will be to derive on general grounds how much can be told simply from symmetry principles. The three light quarks  $u, d$  and  $s$  have mass differences  $\lesssim 150$  MeV, sensibly smaller than the QCD scale<sup>1</sup>. Flavor- $SU(3)$  symmetry results. Even better, the very tiny mass difference  $m_d - m_u$  generates isospin symmetry which is satisfied up to  $\sim 1\%$ . These symmetries are satisfied in the so-called “Wigner mode” i.e. the vacuum is invariant under isospin or flavor- $SU(3)$  rotations and the hadrons are organised in multiplets of  $SU(2)$  (resp  $SU(3)$ ). This reduces the number of unknown parameters. To a good accuracy, it is enough to know the properties of the multiplets rather than those of each individual members of these multiplets. As an example, it is enough to know *one* invariant  $\rho \rightarrow \pi\pi$  amplitude, to know *all*  $\rho^{+0-}$  strong decay amplitudes. Of course these symmetries do not provide any means to compute the invariant amplitude  $\rho \rightarrow \pi\pi$ .

### 3.2 Chiral symmetries

Chiral symmetry is not a symmetry of hadronic states due to spontaneous symmetry breaking. The vacuum is not invariant for chiral symmetry. A chiral rotation generates (annihilates) coherent Goldstone boson states i.e. pions. As a consequence the symmetry tells us almost everything about the soft pion dynamics and yields the effective chiral Lagrangian which fully describes this sector:

---

<sup>1</sup>One can discuss endlessly about what one calls the QCD scale which we shall call generically  $\Lambda_{\text{QCD}}$ . Let us take arbitrarily  $\simeq 1$  GeV, the scale of the nucleon mass.

$$\mathcal{L}^{(2)} = \frac{1}{4} F_\pi^2 \text{Tr}(\partial_\mu U \partial^\mu U^\dagger), \quad (2)$$

where  $U$  is a  $SU(N_f)$  unitary matrix field

$$U = \exp \left( i \frac{\pi_a T_a}{F_\pi} \right), \quad (3)$$

with  $T_a$  the generators of  $SU(N_f)$ , normalized as  $\text{Tr}(T_a T_b) = 2\delta_{ab}$  and  $[T_a, T_b] = 2if_{abc}T_c$  ( $f_{abc}$  being the structure constants of  $SU(N_f)$ ).

As can be seen from (2), all the non-perturbative uncertainty is encoded in the parameter  $F_\pi$  which is furthermore experimentally very well known.

This beautiful situation is unhappily limited to small pion energies. When the latter increases one needs higher terms in the chiral expansion, the number of operators and of unknown constants increases dramatically, not to speak of the renormalisation problems.

### 3.3 Heavy quark effective symmetry

The heavy quark symmetry applies when one quark is heavy (say a  $b$ ) while all other quarks are light (light quarks, gluons, etc). We are then in an atom-like situation in which the heavy quark is almost fixed up to a small recoil motion:

$$p^\mu = m_b v^\mu + k^\mu \quad (4)$$

where  $m_b$  is the heavy mass,  $v_\mu \equiv P_B^\mu/M_B$  the hadron's four-velocity  $k_\mu$  represents the residual motion of the heavy quark relative to the hadron rest frame and  $k^\mu = O(\Lambda_{\text{QCD}})$ .

The heavy quark propagator simplifies:

$$\frac{i m_Q \not{v} + \not{k}}{(m_Q v + k)^2 - m_Q^2} \simeq \frac{\not{v} + 1}{2v \cdot k} \quad (5)$$

as well as the quark-gluon vertex:

$$\bar{u}(s, v) \frac{\not{v} + 1}{2} i t_a \gamma^\mu \frac{\not{v} + 1}{2} i t_b \gamma^\nu \frac{\not{v} + 1}{2} \dots = \bar{u}(s, v) \frac{\not{v} + 1}{2} t_a t_b \dots v^\mu v^\nu \dots \quad (6)$$

This can be done Formally by a change of field variables:

$$h_v(x) \equiv e^{i m_b v^\mu x_\mu} \frac{1 + \not{v}}{2} b(x), \quad H_v(x) \equiv e^{i m_b v^\mu x_\mu} \frac{1 - \not{v}}{2} b(x) \quad (7)$$

and a systematic expansion of the QCD Lagrangian in powers of  $\Lambda_{\text{QCD}}/m_b$  leads to

$$\begin{aligned} \bar{b}(x) (i \not{D} - m_b) b(x) &= \bar{h}_v(x) (i D^\mu v_\mu) h_v(x) + \\ &+ \frac{1}{2m_b} \bar{h}_v(x) (D_\perp^2) h_v(x) + \frac{g_s}{4m_b} \bar{h}_v(x) (\sigma_{\mu\nu} G^{\mu\nu}) h_v(x) \end{aligned} \quad (8)$$

However, life is not exactly so easy. First the heavy quark symmetry assumes that only soft gluons are exchanged by the heavy quark which is only approximately true. Second, the systematics of this two-scale approach needs some rigorous definition of the scale  $\mu$  which separates small from large energy scales. Before considering further this question, let us mention a last effective theory of QCD.

### 3.4 Large energy effective theory

Consider for example a  $B \rightarrow \pi l \nu$  decay. The final pion has a very large momentum in the  $B$  rest frame, and the active final  $u$  quark, which originates from the  $b \rightarrow ul\nu$  decay, has most of the pion energy, the spectator quark being soft. The Large energy effective theory (LEET) [6]-[7] expands the active light quark momentum analogously to the HQET case:

$$p^\mu = En^\mu + k^\mu \quad (9)$$

where  $k^\mu = O(\Lambda_{QCD})$  as in (4),  $E$  is the pion energy in the  $B$  rest frame, i.e.  $E = v \cdot p_\pi$  where  $v_\mu$  is the  $B$  four velocity.  $n^\mu$  is defined so that  $p_\pi^\mu = En^\mu$  with  $v \cdot n = 1$ . Notice that  $n^2 = m_\pi^2/E^2 \simeq 0$ .

There exists a  $SU(2)$  symmetry associated to LEET, but it is not a symmetry of the hadronic states. The reason for that is different than in the case of chiral symmetry: it results from the fact that, even if the active quark takes almost all the momentum of the final pion when it is produced, the strong interaction, after a time  $\sim \Lambda_{QCD}$ , will share the pion momentum between the pion constituents. LEET is then lost, i.e. eq. (9) is no more valid. Still LEET produces very useful relations between  $B \rightarrow \pi(\rho)l\nu$  form factors [7].

## 4 Wilson operator expansion

### 4.1 QCD versus HQET

We have left the HQET with the unanswered question: which is the scale  $\mu$  which separates small/large energy scales ?

Let us consider the example of heavy to light current corrected to one QCD loop.

$$(\bar{q}\gamma_\mu b)_{\text{QCD}} \propto \left\{ 1 + \frac{\alpha_s}{2\pi} \left( \ln \frac{m_b^2}{\lambda^2} - \frac{11}{6} \right) \right\} \gamma_\mu + \frac{2\alpha_s}{2\pi} v_\mu + \dots \quad (10)$$

where  $\lambda$  is an infrared regulator, a “gluon mass”. Notice that eq. (10) depends on the heavy quark mass  $m_b$ . Since the HQET Lagrangian, i.e. the first term in the rhs of (8), does not depend on the heavy mass, the one loop gluon correction in HQET must differ from the one in QCD. Indeed:

$$(\bar{q}\gamma_\mu h_v)_{\text{HQET}} \propto \left\{ 1 + \frac{\alpha_s}{2\pi} \left( \ln \frac{\mu^2}{\lambda^2} + \frac{5}{6} \right) \right\} \gamma_\mu + \dots \quad (11)$$

To reconcile (10) and (11) one expresses the QCD matrix element as the following Wilson [8] operator expansion:

$$(\bar{q}\gamma_\mu b)_{\text{QCD}} = C_1(\mu, m_b) (\bar{q}\gamma_\mu h_v)_{\text{HQET}} + C_2(\mu, m_b) (\bar{q}v_\mu h_v)_{\text{HQET}} \quad (12)$$

with

$$C_1(\mu, m_b) = 1 + \frac{\alpha_s}{\pi} \left( \ln \frac{m_b}{\mu} - \frac{4}{3} \right) + \dots \quad (13)$$

and

$$C_2(\mu, m_b) = \frac{2\alpha_s}{3\pi} + \dots \quad (14)$$

As apparent from (12), the Wilson expansion has separated the two scales in this process: the dependence in large momentum scale,  $m_b$ , is in the coefficients  $C_1, C_2$  while the dependence on the QCD scale is in the HQET operators. While the coefficients can be computed in perturbation theory, the HQET matrix elements need a non-perturbative calculation.

## 4.2 Wilson expansion for inclusive processes

In the preceding subsection we have shown the usefulness of Wilson expansion in exclusive processes, since HQET applies mainly to them, but it has also been fruitfully applied to inclusive  $B$  decays [9]. Let us consider for example the inclusive decay  $B \rightarrow X l \nu$ . As in deep inelastic scattering, we need to compute the imaginary part of the T-product of two currents. It may be Wilson expanded

$$\Gamma(B \rightarrow X l \nu) \propto \int d^4x \operatorname{Im} [ \langle B | T (J_\mu(x) J^\mu(0)) | B \rangle ]$$

whence

$$\Gamma(B \rightarrow X l \nu) = \frac{G_F^2 m_b^5}{192 \pi^3} |V_{CKM}| \times$$

$$c_3(\mu) \frac{\langle B | \bar{b} b(0) | B \rangle_{(\mu)}}{2M_B} + \frac{c_5(\mu)}{m_b^2} \frac{\langle B | \bar{b} \sigma_{\mu\nu} G^{\mu\nu} b(0) | B \rangle_{(\mu)}}{2M_B} + \dots \quad (15)$$

Wilson expansion also has some drawbacks: it introduces order by order a fast increasing number of operators which have to be computed non-perturbatively and often raise renormalisation problems.

## 5 Lattice Monte-Carlo simulations

### 5.1 The principle

We have seen in the preceding sections how symmetries, effective Lagrangians and Wilson expansion can simplify the problem of computing non-perturbative quantities in QCD, reduce the number of non-perturbative unknown and separate the large energy scales, amenable to perturbative treatment, from the QCD energy scale.

But we are still left with the problem of computing from first principles, i.e. from QCD and only QCD, the non perturbative matrix elements of QCD or of the effective theories considered above.

The answer is well known, but unhappily it is only a numerical solution: Lattice Monte-Carlo. The principle is to discretize space and time [10], using lattice spacings typically of the order of  $a \sim 0.03 \rightarrow 0.1$  fm. Numerical calculations also need finite volumes: typically  $L \sim 1.5 \rightarrow 3$  fm.

Last but not least, one performs an analytic continuation to imaginary time (Euclidean metric). It results that the field theory reduces to a Computable 4-D *statistical system* to which standard thermalisation algorithms can be applied [11]. The standard lattice QCD action is [10]

$$S[\mathcal{U}] = - \sum_{x, \mu, \nu} g_{\mu\nu} \frac{2}{g^2} \operatorname{Re} \{ \operatorname{Tr} [1 - P(x)_{\mu, \nu}] \}$$

$$\xrightarrow{a \rightarrow 0} - \frac{1}{4} \sum_{i=1,8} \int d^4x G_{\mu\nu}^i(x) G_i^{\mu\nu}(x)$$

where

$$U_\mu(x) = P \left\{ e^{i a g_0 \int_0^1 d\tau A_\mu^i(x + \tau a \hat{\mu}) \frac{\lambda_i}{2}} \right\}$$

$U_\mu(x) \in SU(3)$  and gauge transformation writes  $U_\mu(x) \rightarrow g(x) U_\mu(x) g^{-1}(x + a \hat{\mu})$ .

Let us skip the methods to introduce quark fields [12]. It is to be noted that the calculation of the quark determinant is very lengthy and computer-time consuming. Therefore the quark determinant is often omitted. This is called the “quenched approximation”. It is equivalent to neglect inner quark loops in Feynman diagrams. It is somehow equivalent to consider only



constituent quarks in hadrons, neglecting sea quarks. This approximation introduces a systematic uncertainty difficult to estimate with full confidence except by comparing with the full lattice QCD calculation, which includes the quark determinant. The latter type of calculation is usually referred to as using “dynamical quarks”.

## 5.2 Some very relevant results

After 25 years of lattice activity the number of results is overwhelming. A huge amount of phenomenological results have been obtained in the quenched approximation. Let us quote some salient results from the section C in the BABAR physics book [13]. These are updated world averages. The heavy meson decay constants are quoted in table 1 to which one should add the phenomenologically crucial result

$$f_B \sqrt{\hat{B}_B^{\text{nlo}}} = 210(42) \text{ MeV}. \quad (16)$$

The renormalisation group invariant  $K - \bar{K}$  mixing parameter:

$$\hat{B}_K = 0.84 \pm 0.07 \pm 0.12 \quad (17)$$

the last error being the quenching error. In table 2 the  $D$  meson decay form factors are given. We do not quote the  $B$  meson decay form factors as the systematic error is still too large. We will return to this shortcoming later.

One must still be aware of the limitations. The masses and momenta in these calculations need to be smaller than the ultraviolet cut-off (the inverse lattice spacing) and larger than the infrared cut-off (the inverse length of the lattice):  $L^{-1} \ll m_q \ll a^{-1}$  i.e.  $50 \text{ MeV} \ll m_q \ll 4 \text{ GeV}$ .

This is why Wilson expansion is extensively used. It is nowadays difficult to implement two different energy scales in the lattice. One computes on the lattice the hadronic matrix elements at a scale of the order of  $\Lambda_{\text{QCD}}$ , and the bridging with the large energy or mass scale is performed in perturbation when computing the Wilson coefficients.

The price to pay is that higher terms in Wilson expansion introduce a large number of operators, the matrix elements of which have to be computed after they have been appropriately renormalised, sometimes a formidable task.

Else one can try extrapolate towards the large energy scale. For example the  $B$  meson decay properties are extrapolated from quarks with a mass of the order of 2 to 3 GeV. This method induces large uncertainties and does not allow to reach large momenta of the decay product.

Another weakness of the present situation is that most of these results are quenched. Some preliminary results using dynamical quarks exist, but they still have large errors and are not yet very useful for phenomenology.

Notwithstanding these limitations, it should be strongly stressed that these are only technical limitations. We mean that, *had we an unlimited computing power and memory space one could predict non-perturbative matrix elements within QCD to any accuracy*, at least as concerns matrix elements with no more than one hadron in the initial and final state: when several hadrons exist together, final state interactions take place, and this makes the analytic continuation back from imaginary time to real time rather difficult.

Let us only mention in passing, although it is a very active and interesting field, the calculations of QCD at finite temperature, as well as the studies of say, chromo-electric flux tubes, and other quantities which give an insight into the confinement mechanism.

## 5.3 Bridging the lattices with perturbative QCD

In section C of [13] one also finds some lattice calculations of fundamental QCD parameters such as quark masses:

$$\bar{m}_s(2\text{GeV}) = 110(2) \text{ MeV from } m_K \quad \bar{m}_s(2\text{GeV}) = 133(6) \text{ MeV from } m_\phi \quad (18)$$

$f_D$	$f_{D_s}$	$f_B$	$f_{B_s}$
$200 \pm 30 \text{ MeV}$	$220 \pm 30 \text{ MeV}$	$170 \pm 35 \text{ MeV}$	$195 \pm 35 \text{ MeV}$

Table 1: Heavy meson decay constants from lattices (world average).

	$D \rightarrow K^{(*)}$ lattice	$D \rightarrow K^{(*)}$ expt	$D \rightarrow \pi, \rho$ lattice
$f^+(0)$	0.73(7)	0.76(3)	0.65(10)
$V(0)$	1.2(2)	1.07(9)	1.1(2)
$A_1(0)$	0.70(7)	0.58(3)	0.65(7)
$A_2(0)$	0.6()	0.45(5)	0.55(10)

Table 2: Heavy meson decay form factors from lattices (world average).

$$\overline{m}_b = 4.15 \pm 0.05 \pm 0.20 \text{ GeV} \quad (19)$$

In [14] and [15] the strong coupling constant of QCD is computed from lattices for zero quark flavors. In [14] the Schrödinger functional is used while in [15] this is done from the Green functions. The lattice spacing is given from the  $\rho$  meson mass, i.e. from a non perturbative quantity, while  $\alpha_s$  is computed at scales up to 5 GeV in [15] or even 100 GeV in [14]. The result converted into  $\overline{\text{MS}}$  scheme for comparison with other estimates,

$$\Lambda_{\overline{\text{MS}}}^{(n_f=0)} = 238(19) \text{ MeV} \quad (20)$$

from Schrödinger functional,

$$\Lambda_{\overline{\text{MS}}}^{(n_f=0)} = 295(5)(15) \text{ MeV} \quad (21)$$

from Green functions.

Figure 1 shows the resulting  $\alpha_s(\mu)$  from [15] compared with the three loops perturbative result. The agreement down to 2.1 GeV is startling. Below that scale non perturbative physics settles in. As expected,  $\alpha_s$  is finite everywhere, i.e. the Landau pole disappears as it is an artifact of perturbative expansion. Here lattices achieve something new and important: they prove to be able to relate perturbative and non-perturbative physics, they confirm deeply the unity of both domains.

Let us now consider the coupling constant for momenta smaller than 2.0 GeV. It is not possible to compare this to well established knowledge. It is however bewildering that the Milan group [16], studying the phenomenology of multi-particle observables find non-perturbative observables which depend on the strong coupling constant in the domain below 2 GeV which turns out to agree quite well with the result in figure 1. More work has to be done in this field but this remark opens new and interesting perspectives.

## 5.4 Some future prospects

The lattice estimates of  $B$  decay parameters such as  $f_B^2 B_B$  and  $B \rightarrow \pi, \rho$  form factors, of  $K - \overline{K}$  mixing, etc.. have been intensively used simultaneously with experimental data to constrain the

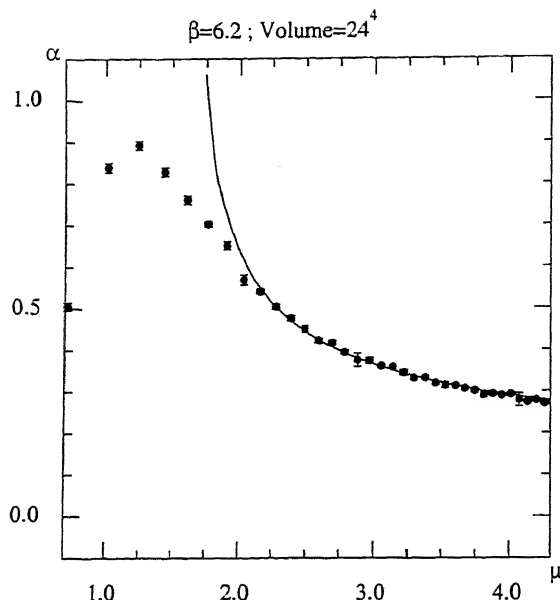


Figure 1:  $\alpha_s$  computed from Green functions on lattices in the  $\overline{\text{MOM}}$  scheme at  $\beta = 6.2$  for a  $24^4$  lattice. The solid curves corresponds to the three loops perturbative alpha computed with  $\Lambda_{\overline{\text{MS}}} = 314$

standard model parameters, and in particular the geometry of the so-called unitarity triangle.

However, the present theoretical accuracy hardly matches the current experimental precision. The latter will quickly improve thanks to the presently starting experiments: BABAR, BELLE, not to speak of Tevatron and the future  $B$  physics at LHC.

Such an improved experimental accuracy will be useless if theory does not match. Beauty phenomenology uses extrapolations up to the  $b$  mass, since the latter overcomes the present ultra-violet cut-offs ( $a^{-1}$ ). As a result some additional uncertainties arise and the accessible  $q^2$  domain is limited. To measure  $V_{ub}$  we need real  $b$  quarks, i.e.  $a \ll 0.04$  fm, i.e.  $\sim 100^4$  lattices if we want to keep a total length of a few fermis. In other words we need to be able to implement directly two scales on the lattice:  $O(m_b)$  and  $O(\Lambda_{\text{QCD}})$ .

On the other hand we would like to reach with *dynamical quarks* an accuracy comparable to today's quenched results.

*This demands  $\sim 10$  teraflops.*

As a bonus, using two scales on the lattice, new attempts could be tried. For example we might also try to compute on the lattice *inclusive* decays, like  $B \rightarrow X l \nu$ ,  $e + N \rightarrow e + X$ , i.e. structure functions, shape functions, etc.

## References

- [1] M Gell-Mann, Phys. Lett. 8 (1964) 214; G. Zweig., CERN preprint TH401 (1964).
- [2] A.N. Mitra and M. Ross Phys. Rev 158 (1967) 1630.
- [3] D.J. Gross and F. Wilczek Phys. Rev Lett 30 (1973) 1343; H.D. Politzer Phys. Rev. Lett. 30 (1973) 1346.
- [4] A.N. Mitra, S. Bhatnagar, I. Santhanam (Delhi U.). Aug 1991. In \*College Park 1991, Hadron'91\* 302-308; A. Le Yaouanc, L. Oliver, O. Pène et J.-C. Raynal, Phys. Lett. 365, (1996) 319.
- [5] For recent reviews see M. Neubert, Lectures given at International School of Subnuclear Physics: 34th Course: Effective Theories and Fundamental Interactions, Erice, Italy, 3-12 Jul

- 1996, hep-ph/9610266; Invited talk at International Europhysics Conference on High-Energy Physics (HEP 97), Jerusalem, Israel, 19-26 Aug 1997, hep-ph/9801269.
- [6] M. J. Dugan and B. Grinstein, Phys. Lett. **B255**, 583 (1991).
- [7] J. Charles, A. Le Yaouanc, L. Oliver, O. Pene, J.C. Raynal, Phys. Rev. **D60** (1999) 014001; Phys. Lett. **B451** (1999) 187.
- [8] K.Wilson Phys. Rev. **179** (1969) 1499; **D3** (1971) 1818; W. Zimmermann, Ann. Phys. **77** (1973) 536 and 570.
- [9] See for example I. Bigi, M. Shifman, N. Uraltsev Ann. Rev. Nucl. Part. Sci. **47** (1997) 591.
- [10] K.G. Wilson, in "New Phenomena in Sub-nuclear Physics", ed. A. Zichichi, Plenum, New york (1977).
- [11] See M. Creutz, Quarks, Gluons and Lattices, (Cambridge Univ. Press 1983) and references therein.
- [12] J. B. Kogut and L. Susskind, Phys. Rev. **D11** (1975) 395; J.L. Alonso, Ph. Boucaud, J.L. Cortés and E. Rivas Mod. Phys. Lett. **A5** 275 (1990). D.B. Kaplan, Phys. Lett. **B288** (1992) 342. For a review see J. Smit Nucl. Phys. B (Proc. Suppl.) **4** (1988) 451.
- [13] THE BABAR PHYSICS BOOK: Physics at an asymmetric B factory: BaBar collaboration. Edited by P.F. Harrison and H.R. Quinn., 1998. (SLAC-R-504)QCD201:B3:1998.
- [14] ALPHA Collaboration (Stefano Capitani et al.) Nucl. Phys. **B544** (1999) 669.
- [15] P. Boucaud, J.P. Leroy, J. Micheli, O. Pene, C. Roiesnel, JHEP **9810** (1998) 017; JHEP **9812** (1998) 004; D. Becirevic, P. Boucaud, J.P. Leroy, J. Micheli, O. Pene, J. Rodriguez-Quintero, C. Roiesnel: hep-ph/9903364.
- [16] Yu.L. Dokshitzer, G. Marchesini, B.R. Webber: hep-ph/9905339.

# 29. QCD Sum Rules in Hadronic and Nuclear Physics

Leonard S. Kisslinger \*

Department of Physics,  
Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A

## Abstract

The Method of QCD sum rules was formulated about two decades ago. It has been a valuable tool for the study of nonperturbative QCD in a large variety of applications. We first review the method as formulated for the treatment of hadronic masses using two-point correlators, with the introduction of the O.P.E and various vacuum condensates. Recent work on glueballs and hybrid hadrons, and the proposed low energy scalar glueball/sigma is discussed. We then review the use of the sum rules for hadronic couplings and form factors, which require three-point functions. Vacuum susceptibilities of the external field method and the use of three-point functions to evaluate them are discussed. The role of nonlocal condensates for form factors and reactions, and light cone sum rules for form factors and wave functions are reviewed. The application of the Dyson-Schwinger formalism for self-consistent studies and its usefulness for sum rule calculations is discussed. The use of instanton solutions for the quark propagator in QCD sum rules is reviewed and the present status of the need for confining guonic effects along with instantons is discussed. Hadrons in nuclear matter and the present status of the attempts to use QCD sum rules to predict mesonic properties in matter at finite temperature as the chiral phase transition is approached and exceeded are reviewed. It is seen that new information on four-quark condensates is necessary for both programs.

## Contents

<b>1</b>	<b>Introduction to Sum Rules</b>	<b>756</b>
1.1	<i>Dispersion Relations</i>	756
1.2	<i>Quantum Chromodynamics</i>	757
1.3	<i>Quark and Gluon Propagators, Vacuum Condensates</i>	758
1.4	<i>QCD Sum Rules</i>	759
1.4.1	<i>Finding Solutions to QCD Sum Rules</i>	760
<b>2</b>	<b>Sum Rules For Masses: Two-Point Functions</b>	<b>761</b>
2.1	Light-Quark Meson and Baryon Masses	761
2.1.1	Meson Masses	761
2.1.2	Factorization of Four-Quark Condensates	763
2.1.3	Heavy-Light Quark Mesons	763
2.1.4	Light-Quark Baryons	763
2.2	Glueballs, Scalar Glueball/Sigma	764
2.2.1	Glueballs	765
2.2.2	Scalar Glueball/Sigma	766
2.3	Mixed Scalar Glueballs and Mesons	766

---

\*Email:kissling+@andrew.cmu.edu

2.4	Hybrid Hadrons . . . . .	767
2.5	Isospin Mass Splittings; Gauge Invariant Electromagnetic Effects; Electromagnetic Penguins . . . . .	768
2.5.1	Isospin Splittings of Charm and Bottom Mesons: Gauge Invariance . . . . .	769
2.5.2	QED Penguins and (B,D) I-spin Mass Splittings . . . . .	770
<b>3</b>	<b>Sum Rules for Coupling Constants: Three-Point Functions</b>	<b>771</b>
3.1	External Field Method: Two-Point Correlator . . . . .	771
3.1.1	Electromagnetic field: Nucleon magnetic Dipole Moments . . . . .	771
3.1.2	Other Fields and Coupling Constants . . . . .	772
3.2	Three-point Function Method for Vacuum Susceptibilities . . . . .	772
3.3	Pion-Nucleon Strong and Weak Coupling . . . . .	774
3.3.1	Pion-Vacuum Correlator for $g_{\pi N}$ . . . . .	774
3.3.2	External Pion Field for $g_{\pi N}$ . . . . .	774
<b>4</b>	<b>Form Factors and Reactions</b>	<b>775</b>
4.1	Form Factors and Nonlocal Condensates . . . . .	775
4.1.1	Pion Wave Function . . . . .	775
4.2	Light-Cone Sum Rules . . . . .	776
4.2.1	Pion Form Factor and Light-Cone Sum Rules . . . . .	776
4.2.2	Pion Wave and Light-Cone Sum Rules . . . . .	777
4.3	Deep Inelastic Scattering; Quark, Sea-Quark Distributions . . . . .	777
4.3.1	Sea-Quark Anisotropy . . . . .	778
<b>5</b>	<b>Dyson-Schwinger, B-S Formalism and Sum Rules</b>	<b>778</b>
<b>6</b>	<b>Instantons and QCD Sum Rules</b>	<b>779</b>
6.1	Instanton Quark Propagator and QCD Sum Rules . . . . .	780
6.2	D-S Study of Instanton Quark Propagator . . . . .	781
<b>7</b>	<b>Hadrons in Nuclear Matter</b>	<b>782</b>
7.1	Nucleons in Nuclear Matter . . . . .	782
7.2	Other Baryons in Nuclear Matter . . . . .	783
7.3	Mesons in Nuclear Matter . . . . .	784
<b>8</b>	<b>Mesons in Finite Temperature Matter</b>	<b>784</b>
<b>9</b>	<b>The Pomeron</b>	<b>787</b>
<b>10</b>	<b>Outlook</b>	<b>788</b>
<b>11</b>	<b>Acknowledgements</b>	<b>789</b>

# 1 Introduction to Sum Rules

The first sum rules were relations associated with the scattering of photons from electrons. These are the Kramers-Kronig[1, 2] dispersion relations, which relate the real to the imaginary parts of the dielectric constant. Such relationships apply to physical scattering amplitudes. Since the imaginary part of a scattering amplitude is related to a total cross section by the optical theorem, a dispersion relation for a scattering amplitude is a sum rule: it relates a scattering amplitude at one energy to well-defined integrals of total cross sections.

In field theories there are dispersion relations for many important quantities, such as propagators and vertex functions. These relations are sum rules. If one can use the field theory to evaluate appropriate quantities in the dispersion relation one has very useful sum rules. In this section we review the general ideas for QCD sum rules.

## 1.1 Dispersion Relations

The dispersion relation for the dielectric constant,  $\epsilon(\omega)$ , where  $\omega$  is the frequency, is derived by recognizing that  $\epsilon$  is analytic in the upper half of the complex  $\omega$  plane and satisfies  $\epsilon(\omega) = \epsilon^*(-\omega^*)$ . It is

$$\begin{aligned} \text{Re}[\epsilon(\omega)] &= 1 + \frac{2}{\pi} P \int_0^{\infty} d\omega' \frac{\omega' \text{Im}[\epsilon(\omega')]}{(\omega')^2 - \omega^2} \\ \text{Im}[\epsilon(\omega)] &= \frac{2\omega}{\pi} P \int_0^{\infty} d\omega' \frac{\text{Re}[\epsilon(\omega')] - 1}{(\omega')^2 - \omega^2}, \end{aligned} \quad (1)$$

where P represents a principal part integral.

For scattering amplitudes, causality ensures the analyticity in the upper half plane and one can derive dispersion relations for scattering amplitudes. One of the early applications of this causal principle for scattering amplitudes[3] showed that for photons scattering from protons, neglecting spin, the forward scattering amplitude,  $f(\omega)$  satisfies the relation

$$\Re[f(\omega)] = -\frac{e^2}{M^2} + \frac{\omega^2}{2\pi^2} \int_0^{\infty} d\omega' \frac{\sigma(\omega')}{(\omega')^2 - \omega^2}, \quad (2)$$

where one has made use of the optical theorem,  $\text{Im}[f(\omega)] = (\omega'/4\pi)\sigma(\omega')$ , with  $\sigma$  is total cross section, and the relation that  $\text{Re}[f(0)] = -e^2/M^2$ .

For field theories the dispersion relation for the forward scattering amplitude is of the form

$$f(\omega) = R(\omega) + \int d\omega' \frac{\omega^n}{(\omega')^n(\omega' - \omega)} \text{Im}[f(\omega')], \quad (3)$$

Where  $n$  is the number of subtractions and  $R(\omega)$  is a polynomial of order  $n$  in the  $\omega$  variable. For example, the forward photon amplitude, Eq(2) is a once-subtracted dispersion relationship. The  $\text{Im}[f]$  is the spectral function. As an example, if there is a stable bound state at energy  $\omega = M_H$  that is in the physical energy region of the scattering the spectral function would be given by a form such as that shown in Fig. 1. Note that in this case at  $M_H$  the spectral function would have a delta function and the forward scattering amplitude a pole.

$$\begin{aligned} \text{Im}[f(E)] &= R\delta(E - M_B) + \text{Im}[f(E)]^{\text{cont}} \\ f(E) &= \frac{R}{\pi(M_B - E)} + \frac{1}{\pi} \int dE' \frac{\text{Im}[f(E')^{\text{cont}}]}{(E' - E)}, \end{aligned} \quad (4)$$

where  $\text{Im}[f^{\text{cont}}]$  is the continuum part of the spectrum.

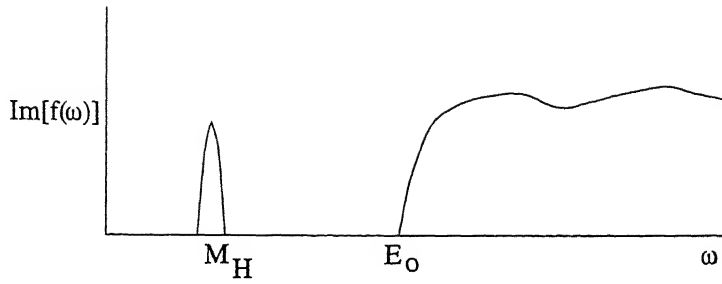
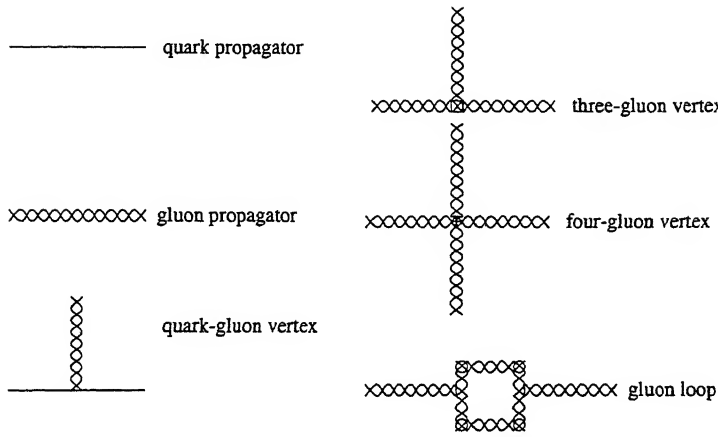
Fig. 1 Spectral function.  $E_0$  is the continuum threshold

Fig. 2 Elements of Feynman Diagrams for QCD

## 1.2 Quantum Chromodynamics

Here we give a very brief review of Quantum Chromodynamics (QCD), a nonabelian gauge field theory. The fields are the quark fields,  $q_f^a(x)$  with color  $a$  and flavor  $f$ , and the gluon field,  $A_\mu(x) = \sum_1^8 A_\mu^c t^c/2$ , where  $t^c$  is an  $SU(3)$  color matrix. The Lagrangian density is

$$\begin{aligned} L^{QCD} &= \bar{q}_f^a (i\gamma^\mu (\partial_\mu - igA_\mu) q_f^a - \frac{1}{2} G^{\alpha\beta} G_{\alpha\beta}) \\ G^{\alpha\beta} &= \partial^\alpha A^\beta - \partial^\beta A^\alpha - ig[A^\alpha, A^\beta]. \end{aligned} \quad (5)$$

The elements of Feynman diagrams used in perturbation treatments of QCD are shown in Fig. 2.

Although QCD is a gauge theory, like QED, it has the property of “antiscreening”, which leads to a number of flavors. In this one-loop calculation the coupling constant becomes infinite at  $q^2 = \Lambda^2$ . QCD at low and medium energies is nonperturbative. This makes the subject most interesting and leads to new fundamental phenomena.

The two systematic ways of studying the structure of hadrons at the present time are numerical lattice gauge calculations and the method of QCD sum rules. In the method of QCD sum rules most of the calculations have used the vacuum condensates in a phenomenological treatment of nonperturbative QCD effects. Another approach to nonperturbative QCD is instanton models, which are consistent with the lowest-dimension condensates, and recently instanton effects have been included in some QCD sum rule calculations, which we discuss in Sec. 6. Next we discuss the nonperturbative quark and gluon propagators.



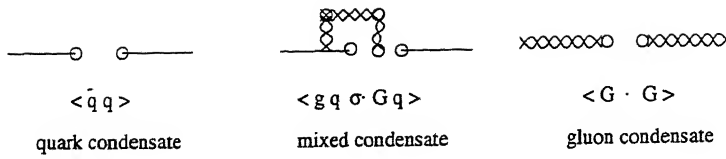


Fig. 3 Vacuum condensates

### 1.3 Quark and Gluon Propagators, Vacuum Condensates

If  $\psi(x)$  is a Dirac field for an elementary particle with mass  $m$ , the perturbative Dirac propagator is given as

$$\begin{aligned} S^{PT}(p) &= \int d^4x e^{ix \cdot p} \langle \bar{0} | T[\psi(x) \bar{\psi}(0)] | \bar{0} \rangle, \\ &= \frac{i}{\not{p} - m}, \end{aligned} \quad (6)$$

where  $T$  is the time-ordering operator,  $\not{p} = \sum_{\mu} \gamma^{\mu} p_{\mu}$  and by  $|\bar{0}\rangle$  we mean the perturbative vacuum, in which all normal ordered products vanish. Using the notation  $|0\rangle$  for the physical, nonperturbative vacuum, the quark propagator is

$$\begin{aligned} S_q(p) &= \int d^4x e^{ix \cdot p} S_q(x) \\ S_q(x) &= \langle 0 | T[q(x) \bar{q}(0)] | 0 \rangle, \end{aligned} \quad (7)$$

where  $q(x)$  is the quark field (with some flavor and color). For the physical vacuum the quark propagator  $S_q(x)$  has a perturbative part,  $S_q^{PT}(x)$ , and a nonperturbative part,  $S_q^{NP}(x)$ ,

$$S_q(x) = S_q^{PT}(x) + S_q^{NP}(x). \quad (8)$$

In the case of vanishing current quark masses ( $m_q = 0$ ) one can write

$$\begin{aligned} S_q^{PT}(x) &= \frac{1}{2\pi^2} \frac{\not{x}}{x^4} \\ S_q^{NP}(x) &= (-) \frac{1}{12} (\langle : \bar{q}(x) q(0) : \rangle + x_{\mu} \langle : \bar{q}(x) \gamma^{\mu} q(0) : \rangle). \end{aligned} \quad (9)$$

It should be stressed that normal-ordered products do not vanish in the nonperturbative vacuum, and therefore and therefore  $S_q^{NP}$  must be considered. For short distances, the O.P.E. for the scalar part of  $S_q^{NP}(x)$  gives

$$\langle : \bar{q}(x) q(0) : \rangle = \langle : \bar{q}(0) q(0) : \rangle - \frac{x^2}{4} \langle 0 | : \bar{q}(0) \sigma \cdot G(0) q(0) : | 0 \rangle + \dots, \quad (10)$$

in which the local operators of the expansion are the quark condensate, the mixed condensate, and so forth. These terms in the quark propagator are illustrated in Fig. 3. The sum rules use Feynman diagrams for the perturbative calculations and Feynman-like diagrams for the nonperturbative processes. Fig. 3 gives the notation for the elements of the nonperturbative processes up to dimension  $D=5$  for use in the Feynman-like diagrams. Since the vacuum condensates are vacuum matrix elements of local operators they are gauge-independent and their values can be determined. Fits to experiments have given the following phenomenological values for the vacuum condensates:  $a = -(2\pi)^2 \langle : \bar{q}(0) q(0) : \rangle \simeq 0.55 \text{ GeV}^3$ ,  $(2\pi)^2 \langle 0 | : \bar{q}(0) \sigma \cdot G(0) q(0) : | 0 \rangle = m_o^2 a$ , with  $m_o^2 \simeq$

$0.8\text{GeV}^2$ . The quark condensate was evaluated using PCAC[4], and the Gell-Mann-Oakes-Renner relation[4],

$$\langle 0|\bar{u}u + \bar{d}d|0 \rangle \simeq \frac{2f_\pi^2 m_\pi^2}{m_u + m_d}, \quad (11)$$

is still an important relation for and consistent with QCD sum rule phenomenology. The systematic treatment of higher-dimension condensates through the operator product expansion[5] is one of the basic elements of the QCD sum rule method, which we discuss in the next subsection.

In a similar way the gluon propagator is defined by the two-point function of the gluon field. The lowest-dimension vacuum condensate for pure glue is the gluon condensate,  $\langle g^2 G \cdot G \rangle = \langle :g^2 G_{\mu\nu}^a G^{a\mu\nu}: \rangle$ . The phenomenological value of the gluon condensate is  $\langle g^2 G \cdot G \rangle \simeq 0.47 \text{ GeV}^4$ .

#### 1.4 QCD Sum Rules

In this subsection we give a brief overview of the QCD sum rule method for obtaining hadronic masses by two-point correlators. Extensions of the method for coupling constants, form factors hadrons in nuclei and so forth will be given in later sections.

A QCD sum rule is a relation between a dispersion relation expression for a correlator and a microscopic QCD evaluation of the correlator. A correlator is defined as a propagator, Eq.(6), in which a composite field operator,  $\eta(x)$ , is used for a hadron in contrast to  $\psi(x)$ , the field function for a basic particle. The composite field operator, usually called a current, must have the property that it can create the hadron under study from the vacuum; i.e., the current  $\eta_H(x)$  for hadron H must satisfy

$$\langle 0|\eta_H|H \rangle = \lambda_H v_H, \quad (12)$$

where  $\lambda_H$  corresponds to the wave function at the origin in a quark/gluon model and, e.g.,  $v_H$  is a Dirac spinor for a spin 1/2 baryon. Clearly, the current does not allow one to obtain all the information about a hadron, as a Bethe-Salpeter amplitude or a wave function in a quantum mechanical system would, but the fact that one starts with local operators for complex systems is an enormous simplification. The correlator is defined by

$$\Pi(p) = \frac{1}{\pi} \int d^4x e^{ix \cdot p} \langle 0|T[\eta(x)\bar{\eta}(0)]|0 \rangle. \quad (13)$$

If the correlator is expressed in terms of a dispersion we refer to it as the phenomenological correlator, since the dispersion relation gives it in terms of physical states and properties. With no subtractions the dispersion relation for the correlator is

$$\Pi(p)^{phen} = \frac{1}{\pi} \int ds \frac{\text{Im}[\Pi(s)]}{s - p^2}. \quad (14)$$

If there is a bound state in the spectrum of hadrons with the quantum numbers of the H of interest, the spectral function will have the form

$$\text{Im}[\Pi_H(s)] = R_H \delta(s - M_H^2) + \text{Im}[\Pi_H(s)]^{cont}, \quad (15)$$

For example let us consider a scalar meson for which  $v_H$  would just be a plane wave in momentum space. Then the form of Eq.(15) leads to the phenomenological expression for the correlation function

$$\Pi(p)_H^{phen} = \frac{1}{\pi} \frac{R^2}{p^2 - M_H^2} + \frac{1}{\pi} \Pi(p)_H^{cont}, \quad (16)$$

where  $\Pi(p)_H^{cont}$  is the continuum contribution to the dispersion relation from  $\text{Im}[\Pi_H(s)]$ .

The QCD calculation is carried out by explicitly using the current  $\eta_H$  in the expression for the correlator, Eq.(13), and calculating the Feynman-like diagrams. We shall see many below. This gives the microscopic QCD correlator,  $\Pi^{QCD}(p^2)$ . One would naively expect the sum rules to arise from

$$\Pi^{phen}(p^2) = \Pi^{QCD}(p^2), \quad (17)$$

but this is not useful, since the QCD series will not converge at low momenta, where the method is being used here. The breakthrough idea of SVZ was to make use of the analytic properties of the correlator to evaluate it at large Euclidean momentum, so that the operator product expansion can be used and rapidly converges. This is done by taking the Borel transform of both sides of Eq.(17). The Borel transform is defined as[5]

$$\tilde{B} = \lim_{Q^2, n \rightarrow \infty, Q^2/n = M_B^2} \frac{(Q^2)^{n+1}}{n!} \left( \frac{d}{d(Q^2)} \right)^n, \quad (18)$$

with  $Q^2 = -p^2$ . A detailed discussion of this transformation and how effective it is for giving convergence is given in Ref. [6]. Note that the Borel transform of a finite polynomial vanishes, so only terms singular at small  $p^2$  contribute. Two important Borel transforms are

$$\begin{aligned} \tilde{B} \frac{1}{(Q^2 + M^2)^r} &= \frac{\exp^{-M^2/M_B^2}}{(r-1)!(M_B^2)^{(k+1)}} \\ \tilde{B}(Q^2)^r \ln(Q^2) &= r!(-M_B^2)^{(r+1)} \end{aligned} \quad (19)$$

Taking the Borel transform of the correlator one arrives at

$$\Pi^{phen}(M_B^2) = \Pi^{QCD}(M_B^2). \quad (20)$$

This is a QCD sum rule. Eq.(19) gives the key to the method. First note that for the phenomenological side the polynomial falloff of the dispersion relation becomes an exponential falloff. This dramatically changes the weight to the states with mass near the Borel mass, which is near the [usually] lowest state being studied, and reduces the effect of the continuum. Referring to Fig. 1 and Eq.(16), note that the part of the spectral function,  $\rho(s) = \text{Im}[\Pi]$ , at  $s = M_H^2 \simeq M_B^2$  becomes more important relative to the continuum part, for which  $s > M_B^2$ . This greatly improves the possibility of finding  $M_H$  accurately in the sum rule. The Borel transform also improves the convergence of the QCD side.

We shall see many examples below, and discuss the techniques used to take into account the spin properties.

#### 1.4.1 Finding Solutions to QCD Sum Rules

Given a current for a hadron, one can always find a sum rule of the form shown in Eq.(20). The question of finding a solution is basic to the method. On the QCD side one calculates the processes up to high enough dimension so that the last calculated term is small, say about 1% of the largest term, and hopes that the higher dimensional terms will not be important. On the phenomenological side one must start with a "good" current; which means that the current for the hadron  $H$  must couple the one- $H$  state to the vacuum so that the structure constant  $\lambda_H$  in Eq.(sr1) is not too small. With this good beginning the main error is in the treatment of the continuum. Often the main uncertainty is the value of the threshold parameter for the continuum, called  $s_0$ . The main criteria for solutions is 1) that one has obtained a region of the Borel mass (the plateau) in which the sum rule Eq.(20) is satisfied with  $M_H$  well within the plateau region, 2) that the continuum contribution is not too large, usually less than 50%, and 3) that the highest dimension terms on the QCD side are very small. This is discussed in many publications. A good discussion of the Borel transformation and its significance in finding solutions is found in Refs.[5, 6].

## 2 Sum Rules For Masses: Two-Point Functions

The study of hadronic masses with the QCD sum rule method is a direct application of the treatment of the two-point correlator outlined in the previous section. One picks an appropriate current for the hadron,  $\eta_H(x)$ , and defines the correlator as in Eq.(13). In this section we review the work on masses with particular emphasis on gluonic hadrons, since these are of greatest interest for current experiments.

### 2.1 Light-Quark Meson and Baryon Masses

The earliest applications of the QCD sum rule methods were for meson and baryon masses, and these have been reviewed in a number of works. Therefore we only give a brief review here. However, the study of mesons in nuclear matter and at finite temperature is of great current interest, and we give some detail for use in later sections. Also, the early work on scalar mesons was not accurate as the mixing with scalar glueballs must be considered. This is discussed in some detail in the next subsection.

#### 2.1.1 Meson Masses

The currents for the mesons are given in Ref.[7], with a detailed discussion of the tensor mesons in Ref.[8]. The currents for the scalar(s) ( $0^{++}$ ), pseudoscalar(ps) ( $0^{-+}$ ), vector(v) ( $1^{--}$ ), and axial vector(av) ( $1^{++}$ ) mesons are of the form

$$\eta^\Gamma(x) = \bar{q}(x)\Gamma q(x), \quad (21)$$

where  $q(x)$  are the u,d-quark fields contracted over their color labels and the  $\Gamma$ 's are the Dirac operators  $\Gamma = 1, \gamma_5, \gamma_\mu$ , and  $\gamma_5\gamma_\mu$ , respectively. The fifth Dirac operator,  $\sigma_{\mu\nu}$ , is not used directly. The axial and tensor currents are more complicated: axial(a)  $\eta_\mu^{(1^{++})} = i\bar{q}\gamma_5\overset{\leftrightarrow}{\partial}_\mu q$ ; tensor(t)  $2^{++}$ ,  $\eta_{\mu\nu}^{(2^{++})} = i\bar{q}(\gamma_\mu\overset{\leftrightarrow}{\partial}_\nu + \gamma_\nu\overset{\leftrightarrow}{\partial}_\mu)q$ ; and the axial tensor(at)  $2^{-+}$ ,  $\eta_{\mu\nu}^{(2^{-+})} = i\bar{q}(\gamma_\mu\gamma_5\overset{\leftrightarrow}{\partial}_\nu + \gamma_\nu\gamma_5\overset{\leftrightarrow}{\partial}_\mu)q$ .

The correlator for each meson,  $\Pi^\Gamma(p)$  is defined in Eq.(13) with a current  $\eta^\Gamma(x)$  as defined above. For the vector, axial vector and axial mesons the correlator is defined as

$$\Pi_{\mu\nu}^{v,av,a}(p) \equiv (p_\mu p_\nu - p^2 g_{\mu\nu})\Pi^{v,av,a}(p); \quad (22)$$

while for the tensor and axial tensor mesons the correlator is defined[8]

$$\Pi_{\mu\nu\alpha\beta}^{t,at}(p) = \frac{(g_{\mu\alpha}g_{\nu\beta} + g_{\mu\beta}g_{\nu\alpha})}{2}\Pi^{t,at}(p). \quad (23)$$

The QCD calculation of the correlators for the light-quark mesons, with the assumption of vanishing quark masses ( $m_q \simeq 0$ ), consists of evaluating three processes shown in Fig. 4, the perturbative diagram (plus a perturbative gluon exchange two-loop diagram not shown), the gluon condensate diagram and the four-quark diagrams.

We now give the contribution of the four-quark diagram for each correlator, since this is of considerable interest in later sections of this review. Let us define  $\hat{Q}^\Gamma$  as the four-quark matrix element for each meson correlator,  $\Pi^\Gamma$ . The  $\hat{Q}^\Gamma$  are defined in terms of the five four-quark condensates

$$\hat{Q}_\Gamma \equiv \langle 0 | : \bar{q}\Gamma t^a q \bar{q}\Gamma t^a q : | 0 \rangle, \quad (24)$$

where  $t^a$  are the SU(3) Gell-Mann color matrices and  $\Gamma$  are the five Dirac operators, and the the four-quark condensate

$$|J_0|^2 \equiv \langle 0 | : \bar{q}\gamma_\mu t^a q \sum_f \bar{q}_f \gamma^\mu t^a q_f : | 0 \rangle. \quad (25)$$

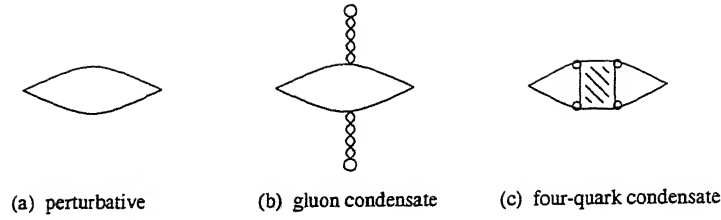


Fig 4. O.P.E. for meson correlator

The  $\hat{Q}^\Gamma$  have the following form:

$$\begin{aligned}
 \hat{Q}^s &\equiv \hat{Q}^t + \frac{2}{3}|J_0|^2, \\
 \hat{Q}^{ps} &\equiv \hat{Q}^{at} + \frac{2}{3}|J_0|^2, \\
 \hat{Q}^v &\equiv \hat{Q}^{av} + \frac{2}{9}|J_0|^2, \\
 \hat{Q}^{av} &\equiv \hat{Q}^v + \frac{2}{9}|J_0|^2, \\
 \hat{Q}^a &\equiv \hat{Q}^{ps} - \frac{1}{9}|J_0|^2, \\
 \hat{Q}^t &\equiv \hat{Q}^v, \\
 \hat{Q}^{at} &\equiv \hat{Q}^{av}.
 \end{aligned} \tag{26}$$

With the phenomenological dispersion relation for the correlator given by Eq(16), one finds the form for the sum rules in  $Q^2 = -p^2$  space

$$\frac{g^\Gamma}{Q^2 + M_\Gamma^2} + \Pi_{cont}^\Gamma(Q^2) = a^\Gamma Q^{2b^\Gamma} \ln(Q^2) + c^\Gamma \frac{\langle g^2 G \cdot G \rangle}{Q^{2d^\Gamma}} + e^\Gamma \frac{\hat{Q}^\Gamma}{Q^{2f^\Gamma}}, \tag{27}$$

with  $g^\Gamma$  a structure constant to be determined and the other constants known. After the Borel transform, with  $M$  the Borel mass, this gives sum rules for the s,v and t mesons

$$\begin{aligned}
 F^s e^{-M_s^2/M^2} &= -\frac{3}{8\pi^2} \left(1 + \frac{11\alpha_s}{3\pi}\right) M^4 - \frac{\alpha_s \langle G \cdot G \rangle}{8\pi} + \frac{\pi\alpha_s}{M^2} \hat{Q}^s + cont. \\
 F^v e^{-M_v^2/M^2} &= \frac{1}{4\pi^2} \left(1 + \frac{\alpha_s}{\pi}\right) M^2 + \frac{\alpha_s \langle G \cdot G \rangle}{12\pi M^2} + \frac{\pi\alpha_s}{M^4} \hat{Q}^v + cont.
 \end{aligned} \tag{28}$$

$$F^t e^{-M_t^2/M^2} = \frac{6}{10\pi^2} \left(1 - \frac{\alpha_s}{\pi}\right) M^6 - \frac{8\alpha_s \langle G \cdot G \rangle}{9\pi} + 4\pi\alpha_s \hat{Q}^t + cont. \tag{29}$$

The ps,av and at sum rules are the same as the s,v and t sum rules with substitution of the superscripts. By dividing the sum rules by their derivatives with respect to  $1/M_B^2$  one can eliminate the unknown phenomenological structure constant,  $F^\Gamma$ . Satisfactory agreement with experiment is found for all the mesons except the ps, which are pseudo goldstone bosons and the scalar mesons, which we discuss below. A recent discussion of methods of obtaining solutions to the sum rules is given in Ref.[9]

### 2.1.2 Factorization of Four-Quark Condensates

In the treatment of meson masses, which we have just discussed, a vacuum saturation hypothesis has been used for the evaluation of the four-quark condensates in order to avoid introducing new parameters. In Ref.[5] this approximation is derived. The factorization assumption gives

$$\langle 0 | : \bar{q}\Gamma t^a q \bar{q}\Gamma t^a q : | 0 \rangle \simeq -\frac{\langle \bar{q}q \rangle^2}{9} (Tr[\Gamma\Gamma] - (Tr[\Gamma])^2) \quad (30)$$

(see the discussion in Sec. 8 on the treatment of mesons at finite temperature), the approximation cannot be used for hadrons in nuclear systems or in finite T matter.

### 2.1.3 Heavy-Light Quark Mesons

The heavy-light (Q-q) quark mesons are the charm-light D mesons and the bottom-light B mesons. In the early days of the use of the QCD sum rule method these mesons were studied[10, 11]. After the concept of heavy quark symmetry was introduced[12] there has been a great deal of work on these systems. For the B systems the masses and decay constants have been calculated using the static heavy-quark approximation and the  $1/m_Q$  expansion[25, 13, 14]; however, for the D systems the charm quark mass is might not be heavy enough for these approximations to be accurate.

The QCD sum rule method was used in a detailed study of the vector and pseudoscalar B and D mesons without using the heavy-quark effective field theory assumptions[15]. The form of the QCD expression, carrying out the O.P.E. to dimension 8, which should give very good accuracy, is

$$\begin{aligned} \Pi(q^2) = & C_1 I + C_3 \langle \bar{q}q \rangle + C_5 \langle \bar{q}(\sigma \cdot G)q \rangle + C_4 \langle \alpha_s G^2 \rangle + \\ & C_6 (g_s \langle \bar{q}q \rangle)^2 + C_7 \langle \alpha_s G^2 \rangle \langle \bar{q}q \rangle. \end{aligned} \quad (31)$$

The coefficients, labelled by the dimension of the local operators, are given in Ref. [15]. The heavy-quark condensates are assumed to be very small and are neglected. Satisfactory values for the vector and pseudoscalar masses were found. The leptonic decay constants, obtained from the structure constants on the phenomenological side of the sum rules, were found to be  $f_B = 95$  MeV,  $f_B^* = 103$  MeV,  $f_D = 130$  MeV and  $f_D^* = 150$  MeV for the ps and v B and D mesons, respectively.

### 2.1.4 Light-Quark Baryons

In the early days of the QCD sum rule method the light-quark baryons were studied[16, 17, 7]. Here we give a very brief review. See Ref.[17] for the sum rules and a detailed description of the results. Details about various aspects of the method are given in Ref.[18]. Since we shall discuss the nucleon and the  $\Delta(1232)$  resonance in nuclear matter in Chapter 7, and discuss the nucleon in the subsection on isospin violations below in this chapter, we briefly review the use of QCD sum rules for these baryons. Applications to the [almost] stable baryon octet and other baryon resonances is very similar for the lightest state with each set of quantum numbers. The current most widely used for the proton is

$$\eta_p(x) = \epsilon^{abc} u^a(x) C \gamma_\mu u^b(x) \gamma^\mu \gamma^5 d^c(x), \quad (32)$$

where the colors a,b,c are summed over and  $C = -C^T$  is the charge conjugation operator. The choice for the proton current is not unique, but arguments have been given[16, 19] for this to be the best current for use in the sum rules. Keeping all processes through dimension nine, the two sum rules obtained as coefficients of  $\not{p}$  and  $I$  are

$$\begin{aligned} \beta^2 e^{-M_p^2/M^2} + \text{continuum} = & \frac{M^6}{8} + \frac{M^2 b}{32} + \frac{a^2}{6} - \frac{a^2 m_o^2}{24M^2} - \frac{m_d a M^2}{4} - \frac{m_d a m_o^2}{24} - \\ & \frac{m_u a m_o^2}{12}, \end{aligned} \quad (33)$$

and

$$\beta^2 M_p e^{-M_p^2/M^2} + \text{continuum} = \frac{aM^4}{4} - \frac{ab}{72} + \frac{34}{81} \frac{\alpha_s a^3}{\pi M^2} + \frac{m_d M^6}{4} - \frac{m_d b M^2}{32} + \frac{m_d a^2}{3} + \frac{m_u a^2}{2}, \quad (34)$$

where  $a \equiv -(2\pi)^2 \langle \bar{q}q \rangle$ ,  $b \equiv \langle g_s^2 G^2 \rangle$ ,  $am_o^2 \equiv (2\pi)^2 \langle g_s \bar{q}\sigma \cdot Gq \rangle$ ,  $\beta^2 = (2\pi)^4 \lambda^2/4$ , with the structure constant defined by  $\langle 0|\eta_p|p \rangle = \lambda v$ , where  $v$  is the proton Dirac spinor. The quantity  $M$  in Eqs.(33,34) is the Borel mass. We have included the quark mass terms, with  $m_q$  the  $u$  and  $d$  current quark masses for the discussion of isospin splitting below, but have not distinguished between the  $u$  and  $d$  quark condensates for reason given below. The continuum is treated by ensuring regularity at large  $M$ , with an important parameter in the fit being  $s_o$ , the threshold  $\text{cm}$  energy squared for the continuum[5]. From these sums rules, corrected to account for the continuum and for the anomalous dimensions of the operators as shown in the  $\Delta(1232)$  sum rules below, the proton mass can be determined to an accuracy of about 10 per cent. The structure constant,  $\lambda$  agrees with lattice gauge calculations to about a factor of two.

The current with no derivatives for the  $\Delta^{++}$  is unique:

$$\eta_\Delta(x)^\mu = \epsilon^{abc} u^a T(x) C \gamma^\mu u^b(x) u^c(x). \quad (35)$$

Defining the correlator with this current as  $\Pi_{\mu\nu}^\Delta$ , one finds two useful sum rules[20] by using  $\text{Tr}[\Pi^\Delta]$  and  $\text{Tr}[\not{p}\Pi^\Delta]$ . They are, respectively,

$$\frac{2}{3} \hat{\lambda}_\Delta^2 e^{-M_\Delta^2/M^2} = \frac{11}{80} M^6 E_2 L^{4/27} - \frac{25}{576} b M^2 E_o L^{4/27} + \frac{5}{6} a^2 L^{28/27} - \frac{35}{72} m_o^2 a^2 L^{16/27} / M^2. \quad (36)$$

and

$$\frac{2}{3} M_\Delta \hat{\lambda}_\Delta^2 e^{-M_\Delta^2/M^2} = \frac{11}{12} a M^4 E_1 L^{16/27} - m_o^2 a M^2 E_o L^{4/27} - \frac{7}{288} ab L^{16/27}. \quad (37)$$

In Eqs.(34,35) the quantities  $a, b, m_o$  and  $M$  are as defined above. The functions  $E_o = 1 - \exp(-x)$ ,  $E_1 = E_o - x \exp(-x)$ ,  $E_2 = E_1 - x^2 \exp(-x)$  are introduced to regulate the large  $M$  behavior of the continuum, with  $x = s_o/M^2$  and  $s_o$  the continuum threshold parameter; and  $L = 0.621 \ln(10M)$  takes into account the anomalous dimensions of the various operators. In the analysis, described in detail in Ref.[20] it was found that  $M_\Delta \simeq (1.35 \pm 10\%)$ .

These are typical examples of the use of QCD sum rules to determine the lightest baryon mass for each set of quantum numbers. Recently there have studies of the negative-parity nucleon resonance[21] and other  $1/2^-$  baryons[22] using the same formalism as in, e.g., Ref.[17]

We now turn to gluonic hadrons.

## 2.2 Glueballs, Scalar Glueball/Sigma

Glueballs, hadrons with gluons as valence particles, are of great interest for the study of the nature of QCD. We expect that the study of glueballs and hybrids, hadrons with both gluons and quarks as valence particles, will dominate hadron spectroscopy in the near future. There have been many theoretical studies of glueballs using the QCD sum rule method[5, 23, 6, 24, 25, 26, 28, 29, 30, 31]. For spectroscopy the scalar glueballs are important not only for their own sake but also because scalar mesons cannot be studied without considering scalar glueballs because of the important scalar meson-glueball mixing[29]. This is a very hot topic, which we discuss in this subsection.

### 2.2.1 Glueballs

The formalism for treating scalar glueballs can be found in Ref.[23] The scalar glueball current is taken to be

$$J_G(x) = \alpha_s G(x) \cdot G(x). \quad (38)$$

Using the once subtracted dispersion relationship, which means that instead of writing the sum rule for the correlator  $\Pi^G$  (Eq.(13) with the current  $J_G$ ), one uses  $(\Pi^G(Q^2) - \Pi^G(0))/Q^2$ . After the Borel transform the sum rule is

$$\begin{aligned} \Pi_G(0)(e^{-M_{GB}^2/M^2} - 1.0) + \frac{2\alpha_s^2}{\pi^2} E_1(s_0) &= 2\frac{\alpha_s^2}{\pi^2} M^4 + 4\alpha_s^2 \langle G \cdot G \rangle \\ &\quad - \frac{8\alpha^2}{M^2} \Gamma^{(6)} - \frac{4\pi\alpha^3}{M^4} \Gamma^{(8)}, \end{aligned} \quad (39)$$

with  $\Gamma^{(6)} = \langle g_s f_{abc} G_{\mu\nu}^a G_{\nu\rho}^b G_{\rho\mu}^c \rangle$  and  $\Gamma^{(8)} = \langle 14(f_{abc} G_{\mu\nu}^a G_{\nu\rho}^b)^2 - (f_{abc} G_{\mu\nu}^a G_{\rho\lambda}^b)^2 \rangle$ . The values of the parameters have been discussed in Refs.[25, 26, 28]. A low energy theorem[6] is used for the phenomenological residue,  $\Pi_G(0) \simeq 3.5 < G^2 >$ .

The remarkable result of the recent calculations of scalar glueballs[25, 26, 28, 27, 29] is the prediction of a light scalar glueball. The greatest uncertainty in the calculation is the value of  $\Gamma^{(6)}$ . In Ref.[29] the sum rule was solved for a range  $-.01 < \Gamma^{(6)} < .05 \text{ GeV}^6$ , with the solutions predicting a scalar glueball in the range  $300 \text{ MeV} < M_{GB} < 600 \text{ MeV}$ . It seems that if one uses the low energy theorem for the gluonic correlator one always finds a good sum rule solution for such a light scalar glueball. No such glueball has been found in lattice gauge calculations[34], however, the dominant nonperturbative effects in the QCD calculations include quark loops that have not yet been included in glueball calculations. We discuss our own conjecture of a scalar glueball/sigma system in the following subsection.

The  $f_0(1500)$ , discovered in the Crystal Barrel experiment[35], is a scalar glueball candidate. From the two-meson branching ratios[36] it is evident that this resonance is not a pure glueball, but is likely to be a mixed scalar glueball-meson system. We return to this from the point of view of QCD sum rules in the subsection below. It should also be noted that the dominant decay of the  $f_0(1500)$  into four pions (two sigmas). We return to this in the next subsection of the scalar glueball/sigma.

The tensor glueball is also of great current interest. The  $\xi(2230)$  discovered in the MARK III[37] and BES collaboration[38] experiments is a candidate for a tensor glueball, although its spin-parity has not yet been determined. It also shows a strong four-pion (two-sigma) decay, which might be a signal for glueballs, as we discuss below. The gauge invariant current for the tensor glueball which has been used in QCD sum rule calculations is[6]

$$J_{\mu\nu}^{G(2^{++})} = -G_{\nu}^{\alpha\alpha} G_{\nu\alpha}^a + \frac{1}{4} g_{\mu\nu} G^{\alpha\alpha\beta} G_{\alpha\beta}^a. \quad (40)$$

Solutions using the QCD sum rule method are found in the range of 2.0-2.7 GeV. These are values found in lattice gauge calculations. See Ref.[31] for a recent review of calculations of the tensor glueball.

An important consideration for the study of glueballs is the mixing with mesons, and this is particularly important for the scalar case, where there is a low energy theorem[6] that is most important for sum rule calculations, as we shall discuss below. Estimates of the scalar glueball/meson mixing[32] and of the tensor glueball-meson mixing[33], without the use of the low-energy theorems of Ref.[6] have found very small mixing.



### 2.2.2 Scalar Glueball/Sigma

The sigma has been a part of nuclear physics for decades in various forms. When it was found that meson exchange potentials using the known mesons could not produce enough attraction to explain the nuclear force, a scalar meson, the  $\sigma$ , was introduced. It often has a mass range of 600-800 MeV. In effective chiral field theories the  $\sigma$  is introduced as a chiral partner to the pion. More recently, the sigma has been identified as a  $\pi$ - $\pi$  low-energy resonance. That is what we refer to as the sigma in the present paper.

In a K-matrix analysis of the  $I=0, L=0$  ( $0^{++}$ ) channel of the  $\pi - \pi$  scattering amplitude[39] a low-lying resonance was found. In this analysis the  $f_0(980)$  and higher resonances were subtracted out. The remaining amplitude shows an almost elastic Breit-Wigner form:

$$A_{\sigma}^{\pi\pi(L=0)} = \frac{-M_{\sigma}\Gamma_{\sigma}}{s - M_{\sigma}^2 + iM_{\sigma}\Gamma_{\sigma}}, \quad (41)$$

with  $M_{\sigma} \simeq \Gamma_{\sigma} \simeq 400$  MeV. This resonance is what we refer to as the sigma.

Moreover, in an analysis of the dominant  $4\pi$  branching ratios of all of the current candidates for glueballs the BES group found[38] that this decay channel is completely dominated by the two- $\sigma$  mechanism. GLUE IS STRONGLY COUPLED TO THE SIGMA! Furthermore, the mass of the sigma is in the range of our scalar glueball solution.

From these observations we have made our scalar glueball/sigma ansatz: The sigma phenomenon is a two-channel glueball, two- $\pi$  system. With this ansatz we can extract the sigma-glueball coupling constant:

$$g_{\sigma} \simeq \Gamma_{\sigma} \simeq 360 \text{ MeV} \quad (42)$$

If this picture is valid it will provide a signal for gluonic hadrons.

### 2.3 Mixed Scalar Glueballs and Mesons

It was pointed out in the early study of scalar glueballs[6] that there is a low-energy theorem for the coupling of the scalar glueball and meson currents

$$\int d^4x \langle 0 | T(J_G(x) J_m(0)) | 0 \rangle = \frac{32}{9} \langle \bar{q}q \rangle. \quad (43)$$

This indicates that there might be important mixing between the scalar glueballs and mesons. Indeed this is the case. In Ref.[29] the current for a mixed scalar meson-glueball

$$J_{0^{++}} = \beta M_{\sigma} J_m + (1 - |\beta|) J_G, \quad (44)$$

with  $M_{\sigma} = 1$  GeV, was used. This leads to a QCD sum rule with a microscopic side after the Borel transform

$$\Pi_{QCD}^{0^{++}GB-m}(M) = \beta^2 \Pi_{QCD}^{0^{++}m}(M) + (1 - |\beta|)^2 \Pi_{QCD}^{0^{++}GB}(M) + \frac{64}{9} \beta (1 - |\beta|) \langle \bar{q}q \rangle. \quad (45)$$

Solving the sum rule for  $\beta = 0$  gives the light scalar glueball solution discussed above. The unmixed meson solution found with  $\beta = 1$  was shown[40] to give a solution about with a mass about 1 GeV, which was interpreted as a successful fit to the  $f_0(980)$ . Most people working in meson spectroscopy now believe that the  $f_0(980)$  is not a meson (meaning a  $\bar{q}q$  state, but is more likely to be a more complicated state.

The solutions found in Ref[29] were an almost meson solution with  $|\beta| \simeq 0.8$  at about 1400 MeV, corresponding to the  $f_0(1370)$  meson-like resonance and an almost glueball solution with  $|\beta| \simeq 0.2$  corresponding to  $f_0(1500)$  found in the Crystal Barrel experiment[35]. Therefore we see

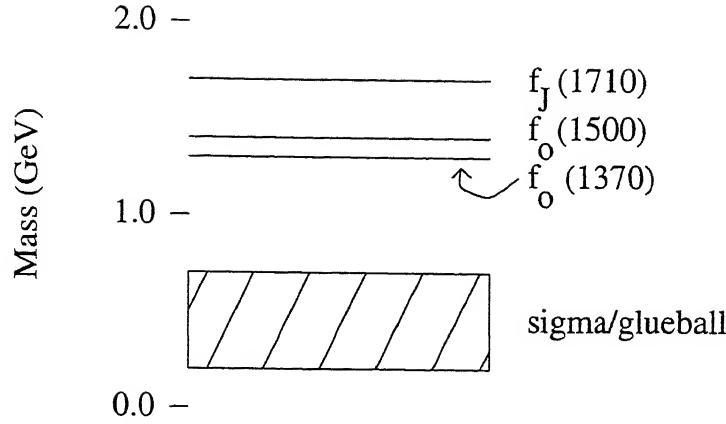


Fig. 5 Scalar meson/glueball states

that in the channel with vacuum quantum numbers ( $0^{++}$ ) meson-glueball mixing must be treated for an accurate theoretical description.

The summary of the QCD sum rule studies of scalar hadrons at the present time is shown in Fig.5. The  $f_0(980)$  is not shown, as discussed above. There are two established  $f_0$  hadrons, the  $f_0(1370)$  and the  $f_0(1500)$ , which we have just discussed. If the  $f_J(1710)$  proves to be another  $f_0$  (i.e.,  $J=0$ ) then it is likely that the three  $f_0$  states are admixtures of light-quark, strange quark and glueball scalar systems, which would be consistent with lattice gauge calculations[34]. Such admixtures have been discussed in Ref[31]. It is now most important to seek experimental signals for such states.

## 2.4 Hybrid Hadrons

Hybrid hadrons have valence glue. The QCD sum rule method is an excellent tool for studying hybrids as they can be quite precisely defined, which is not the case for many quark models. For example, a current of the form

$$\eta^G = \epsilon^{abc} q^a \Gamma_1(\bar{q})^b \Gamma_2 G^c \quad (46)$$

is a hybrid meson, since the quark-anticolor system carries net color which combines with the valence glue to give a color = 0 hybrid. Mesons are excellent systems for hunting hybrids since there are exotic quantum numbers which cannot exist in  $q\bar{q}$  systems. An example is  $J^{PC} = 1^{-+}$ . Early in the QCD sum rule days calculations of the  $1^{-+}$  were carried out[41, 42, 43], with the authors predicting this exotic state at about 1.5 GeV. Recently the Crystal Ball group has reported a  $1^{-+}$  state at about 1.4 GeV[44] seen in the  $\eta - \pi$  channel, in nice agreement with theory. Flux-tube models predict[45, 46] the state at about 1.8 GeV. The nature of the state as a four-quark or hybrid is not certain, and a study of the  $\eta - \pi$  decay[47] gives valuable information about the nature of this exotic system.

Since baryons do not have exotic quantum numbers, it is more difficult to determine that one has found a hybrid baryon. There have been a number of QCD sum rule calculations for the  $J^P = 1/2^+$  hybrid, with the same quantum numbers as the nucleon[48, 49, 50]. In Ref.[50] the current used for the hybrid was

$$\eta_H(x) = \epsilon^{abc} (u^a(x)^T C \gamma^\mu u^b(x)) i \sigma^{\alpha\beta} \gamma^\mu \gamma^5 G_{\alpha\beta}^d(x) (t^d d(x))^c, \quad (47)$$

which is conveniently renormalized. Note that the form of Eq.(47) with the properties of the SU(3) color generator,  $t^d$  ensures that the three quarks are not in a color zero configuration and that the gluon is a valence particle.

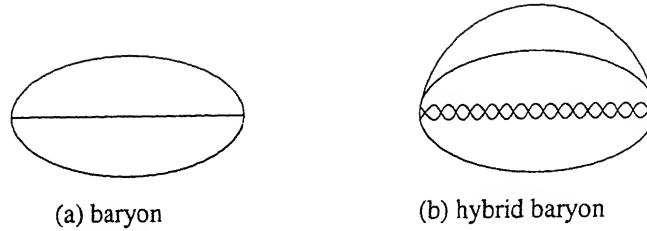


Fig. 6 Lowest dimension correlator for baryon and hybrid baryon

The correlator of the nucleon and the hybrid are illustrated in Fig.6. A very stable solution was found at  $M_H \sim 500$  MeV above the nucleon mass. In other words it is predicted that the  $P_{11}(1440)$ , the so-called Roper resonance, is a hybrid, or at least that there is a hybrid at that mass. Subsequently [30], it was shown that the nucleon has very little hybrid admixture. Using the idea of the glueball/sigma discussed above, it was shown[51] that the decay of the Roper into a nucleon and a two-pi (sigma) could be a test of the hybrid nature of this baryon. The same applies to the decays of many hybrid hadrons.

## 2.5 Isospin Mass Splittings; Gauge Invariant Electromagnetic Effects; Electromagnetic Penguins

In QCD the only sources of isospin violations are the quark mass differences, and the mass splittings of hadronic isospin multiplets in principle give important information about the current quark masses in the QCD Lagrangian. In quark models the sources of isospin mass splittings are electromagnetic effects and the constituent quark mass differences. For example, for the neutron-proton mass splitting the d quark must be heavier than the u quark since electromagnetic effects would make the proton heavier than the neutron. In the QCD sum rule method it is much less straightforward. As one can see from the QCD correlator for the proton, shown in Eqs.(33,34) there are three sources of isospin splitting

$$\begin{aligned}
 \text{quark mass difference} &= m_d - m_u \\
 \text{condensate difference} &= \langle \bar{u}u \rangle - \langle \bar{d}d \rangle \\
 \text{electromagnetic effects} &
 \end{aligned}
 \tag{48}$$

In a study of the neutron-proton mass differences Eqs.(33,34) were modified to treat the resonances, the continuum and the anomalous dimensions, and the I-spin splittings of the quark and mixed condensates[52]. Although it was not pointed out in that paper, one can show that if one neglects electromagnetic and condensate effects the sum rules predict

$$m_d > m_u \implies M_p > M_n. \tag{49}$$

This is certainly counter-intuitive, since one expects that if the neutron with two d quarks and one u quark would be more massive than the proton with one d quark and two u quarks if the d-quark is more massive than the u-quark. This implies that the I-spin splittings of the condensates, which in turn must arise from the d-u mass differences, give a larger effect than the quark mass difference itself. Since we are dealing with nonperturbative phenomena, we can expect such surprises. We shall not give a review of the QCD sum rule calculations of the baryon I-spin mass differences, since there is a problem in all of the calculations, including our own in treating the electromagnetic corrections. We now discuss this.

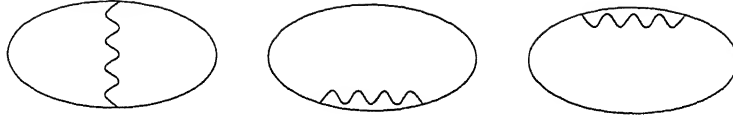


Fig. 7 Photon loops, usual QED diagrams

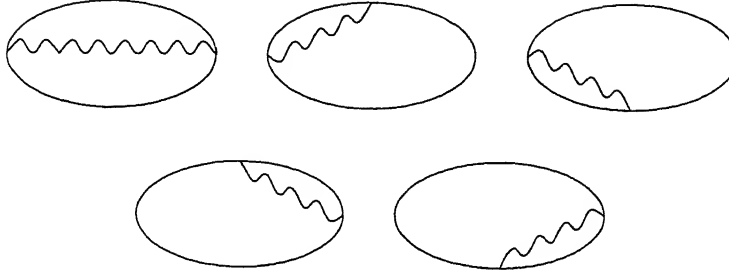


Fig. 8 Additional photon loop diagrams for electrically charged meson

### 2.5.1 Isospin Splittings of Charm and Bottom Mesons: Gauge Invariance

The heavy-light quark mesons are excellent systems in which to study isospin violations since the nonperturbative QCD effects are smaller than in light-quark systems. As discussed above the QCD sum rule method gives satisfactory agreement with experiment for masses of the vector and pseudoscalar B and D mesons. For calculations of isospin mass splittings one must include the effects shown in Eq.(48). This was done in a calculation[53] of B(ps), B\*(vector), D and D\* mass splittings. The startling thing was that the standard two-loop photon diagrams expected to give the correct QED effects to order  $\alpha_{QED}$ , shown in Fig. 7, are not gauge invariant. I.e., if one makes a gauge transformation

$$\begin{aligned} A_\mu(x) &\Rightarrow A_\mu(x) - \partial_\mu \Lambda(x) \\ \bar{q}(x) &\Rightarrow e^{ie_q \Lambda(x)} \bar{q}(x) \\ Q(x) &\Rightarrow e^{ie_Q \Lambda(x)} Q(x), \end{aligned} \quad (50)$$

where  $q$  and  $Q$  are the light-quark and heavy-quark fields,  $A_\mu$  is the electromagnetic field and  $\Lambda$  is the gauge function, then the correlator (say for a scalar meson)

$$\Pi(q^2) = i \int d^4x e^{iqx} \langle T[J(x)J(0)] \rangle, \quad (51)$$

with  $J(x) = \bar{q}(x)Q(x)$ , is not gauge invariant. In order to obtain a gauge invariant solution, consider the correlator with the QED link operator

$$\Pi(q^2) = i \int d^4x e^{iqx} \langle T[J(x) \exp[iQ_{op} \int_y^x dy \cdot A(y)] J(y)] \rangle, \quad (52)$$

where  $Q_{op}$  is the charge operator. By making an expansion in  $A_\mu$  it was shown[54] that the additional vertex diagrams shown in Fig. 8 must be included, and that with all of the two loop diagrams shown in Figs.7,8 the calculation is gauge invariant. If one uses the quark mass difference  $m_d - m_u = 3.8$  MeV, estimated from broken chiral symmetry[55], and takes values for the parameter

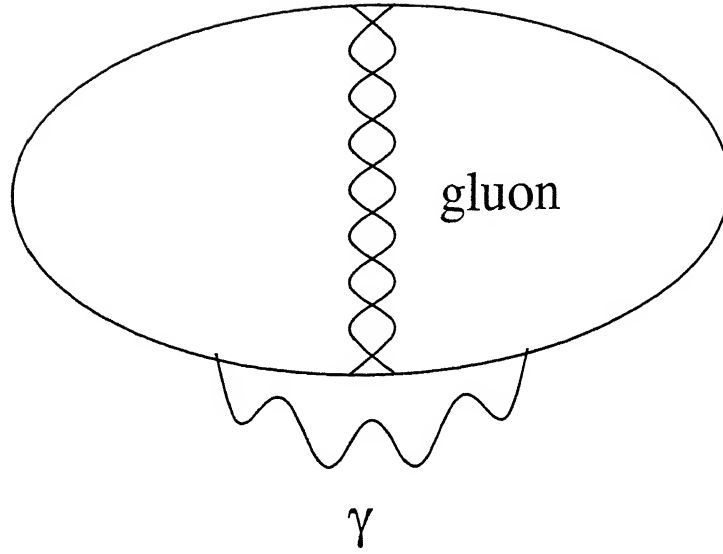


Fig.9 QED Penguin diagram

$\gamma$  for the quark condensate isospin splitting

$$\gamma = \frac{\langle \bar{d}d \rangle}{\langle \bar{u}u \rangle} - 1, \quad (53)$$

in the range  $-.002 \rightarrow -.0079$ , as estimated in fits to the neutron-proton mass splitting[52, 56] then one finds[54]

$$\begin{aligned} M_D^\pm - M_D^0 &= 4.3 \rightarrow 4.7 \text{ Mev} \\ M_D^{\pm*} - M_D^{0*} &= 2.6 \rightarrow 3.0 \text{ Mev} \\ M_B^\pm - M_B^0 &= -3.2 \rightarrow -3.0 \text{ Mev} \\ M_B^{\pm*} - M_B^{0*} &= -3.6 \rightarrow -3.5 \text{ Mev} \end{aligned} \quad (54)$$

for the pseudoscalar and vector, D and B splittings, resp. These results are quite satisfactory for the charmed (D) mesons, but the mass splittings for the bottom mesons is much too large in comparison with experiment[57]. As expected, the nonperturbative quark condensate I-spin violation effect is small, so to the extent that the QED effects have been taken care of correctly, these mass splittings should be a good source on information on the current quark mass splittings. Clearly there is a problem with the B mass splittings. We believe that the source of the error is a missing process, the QED penguin diagrams. We now discuss this.

### 2.5.2 QED Penguins and (B,D) I-spin Mass Splittings

Several years ago it was pointed out that there is a QED effect that modifies the quark-gluon vertex in a manner analogous to the so-called penguin modification of the quark gluon by the weak fields[58]. For the heavy-light mesons these QED penguin modifications are illustrated in Fig. 9, in which the penguin modification of the two-loop perturbative gluon is illustrated. This diagram and penguins for the mixed quark condensate processes were included in the calculation of Ref.[54] in Ref.[59]. It was shown that with the penguins the theory is consistent with the very small B isospin splitting, but there were inaccuracies in the calculation. The effects in the D systems were small.

### 3 Sum Rules for Coupling Constants: Three-Point Functions

The coupling of a field to a hadron involves a three-point function. Consider a current  $J^\Gamma(y) = \bar{q}(y)\Gamma q(y)$  coupling to a hadron  $\alpha$  with a composite field operator  $\eta_\alpha(x)$  to form another hadron  $\beta$  with a composite field operator  $\eta_\beta(x)$ . In field theory this is treated by the three-point function

$$V_{\beta\alpha}^\Gamma(p, q) = \int d^4x \int d^4y e^{ix \cdot p} e^{-iy \cdot q} \langle 0 | T[\eta_\beta(x) J^\Gamma(y) \bar{\eta}_\alpha(0)] | 0 \rangle \quad (55)$$

The treatment of three-point functions is much more complicated than two-point functions. Moreover, for coupling constants, in which there is very little or zero momentum transfer at the vertex, the O.P.E. used for the two-point correlator via the Borel transform cannot be used directly as the point  $y$  in Eq.(55) is not at short distance from points  $x$  or  $0$  for small  $q$ . To solve this problem the external field method was introduced into QCD sum rules, which we now discuss.

#### 3.1 External Field Method: Two-Point Correlator

In the study of the magnetic dipole moments of nucleons the problem of avoiding the O.P.E. in the  $y$  variable was discussed at length[60] in a three-point calculation using the general ideas of Ref.[5]. To avoid these difficulties a two-point correlator for a nucleon in an external field was introduced[61] for the case of electromagnetism. For an external current  $J^\Gamma$  coupled to a proton the correlator is

$$\Pi^\Gamma(p) = i \int d^4x e^{ix \cdot p} \langle 0 | T[\eta_p(x) \bar{\eta}_p(0)] | 0 \rangle_{J^\Gamma} \quad (56)$$

where  $\eta_p$  is the nucleon current given in Eq.(32). As can be seen from Eq. (56) the microscopic evaluation of  $\Pi^\Gamma(p)$  can be done using the operator product expansion, since the variable  $x$  is at short distance from the origin. This is done by an O.P.E. of the quark propagator in the presence of the  $J^\Gamma$  current

$$\begin{aligned} S_q^\Gamma(x) &= \langle 0 | T[q(x) \bar{q}(0)] | 0 \rangle_{J^\Gamma}, \\ &= S_q^{\Gamma, PT}(x) + S_q^{\Gamma, NP}(x), \end{aligned} \quad (57)$$

where  $S_q^{\Gamma, PT}(x)$  is the quark propagator coupled perturbatively to the current and  $S_q^{\Gamma, NP}(x)$  is the nonperturbative quark propagator in the presence of the external current,  $J^\Gamma$ .

For the two-point treatment at low momentum transfer the O.P.E. for  $S_q^{\Gamma, NP}(x)$  is justified as in the ordinary two-point function, giving

$$S_q^{\Gamma, NP}(x) = \frac{-\Gamma}{12} \langle 0 | : \bar{q} \Gamma q : | 0 \rangle_{J^\Gamma} + \frac{x^2 \Gamma}{3 \cdot 2^6} \langle 0 | : \bar{q} \sigma \cdot G \Gamma q : | 0 \rangle_{J^\Gamma} + \dots \quad (58)$$

This is illustrated in Fig. 10. This method has been applied to the couplings of a number of fields to the nucleon.

##### 3.1.1 Electromagnetic field: Nucleon magnetic Dipole Moments

For the electromagnetic field the field-quark coupling is given by the quark charge,  $e_q$  as shown in Fig. 10(a). The field polarizes the quark condensate, as shown in Fig. 10(b). The corresponding nonperturbative constant, called the magnetic quark susceptibility,  $\chi$ , is defined by

$$\langle \bar{q} \sigma_{\mu\nu} q \rangle = e_q \chi F_{\mu\nu} \langle \bar{q} q \rangle, \quad (59)$$

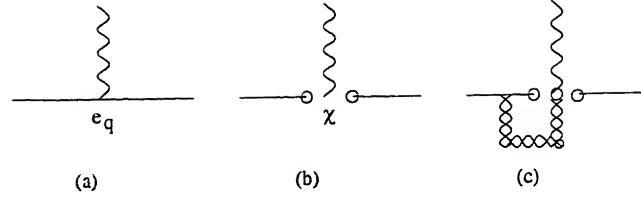


Fig. 10 (a) field-quark coupling (b) magnetic susceptibility (c) mixed susceptibility

where  $F_{\mu\nu}$  is the QED field tensor. Two other susceptibilities are associated with the polarization of the mixed quark condensate[61], illustrated in Fig. 10(c).

There are three covariants in the expression for the correlator in the external electromagnetic field, so one obtains three sum rules for the proton and neutron, two of which are used. By manipulating these equations one can eliminate the susceptibilities and obtain approximate expressions for the neutron and proton moments in terms of the quark condensate:

$$\begin{aligned}\mu_p &\simeq \frac{8}{3}\left(1 + \frac{a}{6M_N^3}\right) \\ \mu_n &\simeq -\frac{4}{3}\left(1 + \frac{2a}{3M_N^3}\right),\end{aligned}\quad (60)$$

where  $a \equiv -(2\pi)^2 \langle \bar{q}q \rangle \sim .55 GeV^3$ . With an estimated 10% errors the method is in agreement with the experimental values of  $\mu_p^{exp} = 2.79 nm$ ,  $\mu_n^{exp} = -1.91 nm$ . Estimates of the magnetic quark susceptibility give

$$\chi a M_p \simeq -(3.4 \rightarrow 6.0) GeV^2. \quad (61)$$

Estimates of  $\chi$  have been made in Refs.[60, 62] using rho-meson dominance models. We shall return to the determination of vacuum condensate susceptibilities using a three-point method below.

### 3.1.2 Other Fields and Coupling Constants

The external field method has been used for the calculation of the coupling constants for a number of field-nucleon vertices. The method is similar to the calculation of the nucleon magnetic dipole moments reviewed in the previous subsection, so we shall not give details. There have been a number of applications of this two-point method for the calculation of the axial coupling constant ( $g_A$ ) [63, 64, 65] and for the isoscalar axial coupling constant ( $g_A^S$ )[65]. From the appropriate sum rules the axial vector susceptibility has been estimated. There as also been an estimate of the vacuum tensor susceptibility[66] in work on the tensor charge of the nucleon, however, it has been pointed out[67] that the treatment of the nucleon tensor charge is subtle and that different theoretical treatments can give very different results. In all of these calculations there are new unknown susceptibilities. They are treated in a variety of ways, often manipulating the sum rule equations to eliminate them. We shall show that in an approximation to the basic three-point method there is an almost universal value for these susceptibilities. We shall discuss the strong and weak pion-nucleon coupling separately

## 3.2 Three-point Function Method for Vacuum Susceptibilities

Let us now return to the three-point function formulation, Eq.(55), for the current  $\Gamma$  coupling to the nucleon,

$$\begin{aligned}V^\Gamma(p, q) &= \int d^4x \int d^4y e^{ix \cdot p} e^{-iy \cdot q} V^\Gamma(x, y) \\ V^\Gamma(x, y) &= \langle 0 | T[\eta(x) J^\Gamma(y) \bar{\eta}(0)] | 0 \rangle,\end{aligned}\quad (62)$$

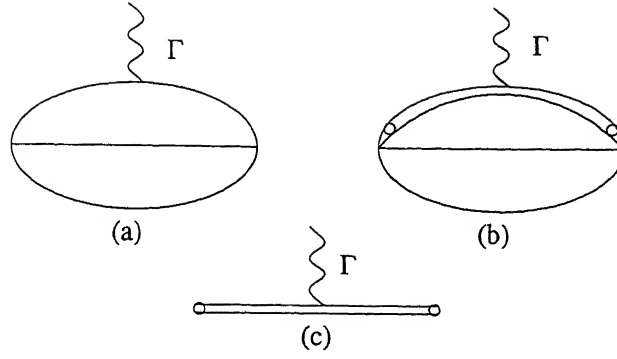


Fig. 11  $\Gamma$ -current (a) coupling, (b) susceptibility, (c)  $S_q^{\Gamma, NP}$

with  $\eta$  given in Eq.(32). Note that  $V^\Gamma(x, y)$  has terms with vacuum matrix elements involving two, four, six and eight quark fields. It is the four-quark term that is of interest to us for the study of the vacuum susceptibilities. This is illustrated in Fig.11 The process of Fig.11(a) is the coupling of the field to the perturbative quark, while the process of Fig.11(b) corresponds to the polarization of the quark condensate by the  $\Gamma$  field in the external field method. In Ref.[68] the pion current with the form  $\Gamma = \gamma_5$  was considered. The four-quark term of the three-point function corresponding to this pion current is

$$V^{\Gamma 4q}(x, y) = -i2\epsilon^{abc}\epsilon^{b'a'c'} \langle 0 | \gamma^5 \gamma_\mu d^c(x) \bar{d}^e(y) \Gamma d^e(y) d^{c'}(0) \gamma_\nu \gamma^5 | 0 \rangle \\ Tr[S_u^{aa'}(x) \gamma^\mu C(S_u^{bb'}(x))^T C \gamma^\nu]. \quad (63)$$

The quark  $\Gamma$  susceptibility, corresponding to Fig11(c) is given by

$$S_q^{cc'\Gamma, NP}(x) = -i \int d^4y \langle 0 | : q^c(x) \bar{q}^e(y) \Gamma q^e(y) \bar{q}^{c'}(0) : | 0 \rangle \quad (64)$$

in the limit of  $q_\mu \rightarrow 0$ . The expression for  $S_q^{cc'\Gamma, NP}(x)$  is evaluated using factorization and a nonlocal quark condensate

$$\langle 0 | : \bar{q}(0) q(y) : | 0 \rangle \equiv g(y^2) \langle 0 | : \bar{q}(0) q(0) : | 0 \rangle. \quad (65)$$

The form for the quark condensate nonlocality in Ref.[68] was taken to be

$$g(y^2) = \frac{1}{(1 + \kappa^2 y^2/8)^2}, \quad (66)$$

with the parameter  $\kappa$  evaluated using the sea-quark distributions obtained with the QCD sum rule method of deep inelastic scattering of Ref.[69]. The resulting expression for the nonperturbative quark propagator in the  $\Gamma$  field is

$$S_q^{\Gamma, NP}(x) \simeq \Gamma G(x) \langle 0 | : \bar{q}(0) q(0) : | 0 \rangle / 12^2, \\ G(x) = (-i) \int d^4y g(y^2) g((x-y)^2). \quad (67)$$

For the case of the pion susceptibility

$$\langle 0 | : \bar{q} i \tau_3 \gamma^5 q : | 0 \rangle_\pi = \chi^\pi \langle 0 | : \bar{q} q : | 0 \rangle, \quad (68)$$

one finds that

$$\chi^\pi a \simeq -\frac{2a^2}{9\kappa^4} \\ \simeq -(1.7 - 3.0) GeV^2, \quad (69)$$



in agreement with the value  $\chi^\pi \simeq -1.88 \text{ GeV}^2$ , found in Ref.[70], which we discuss in the next subsection. Here we have changed the sign of  $\chi^\pi$  from that used in Ref.[68] in order to agree with the sign choice in Ref.[70]. Note that this value is more than an order of magnitude different from the value one obtains using a two-point method with PCAC[71]. There has also been an estimate of  $\chi^\pi$  using effective chiral quark theory[72].

In subsequent work[73] this formalism was shown to be in agreement with all known susceptibilities. Therefore the expression of Eq.(67) is an approximate general relationship for all susceptibilities.

### 3.3 Pion-Nucleon Strong and Weak Coupling

The pion-nucleon coupling constant was determined many years ago from an extrapolation of  $\pi - N$  scattering, as well as two-nucleon properties:  $g_{\pi N} \simeq 13.4$ . It has been estimated using QCD sum rules using two different two-point methods.

#### 3.3.1 Pion-Vacuum Correlator for $g_{\pi N}$

Before the external field method was published,  $g_{\pi N}$  was estimated using a two-point correlator but with a low-energy pion in the initial state [7] rather than an external pion field as in Eq.(57).

$$\Pi_1^\pi(p) = i \int d^4x e^{ix \cdot p} \langle 0 | T[\eta_p(x) \bar{\eta}_p(0)] | \pi \rangle \quad (70)$$

Later a much more detailed calculation of  $g_{\pi N}$  was carried out using this method[75]. Since there is no external field, in this method there is no pion susceptibility. The pion quark vacuum susceptibility is replaced by the matrix element

$$\langle 0 | i \bar{q} \gamma^5 q | \pi \rangle = \frac{\langle \bar{q} q \rangle}{f_\pi}, \quad (71)$$

which has been evaluated using a soft pion theorem, where  $f_\pi$  is the pion decay constant. In Ref.[75] the difficulty in getting a reliable value for  $g_{\pi N}$  is discussed in detail.

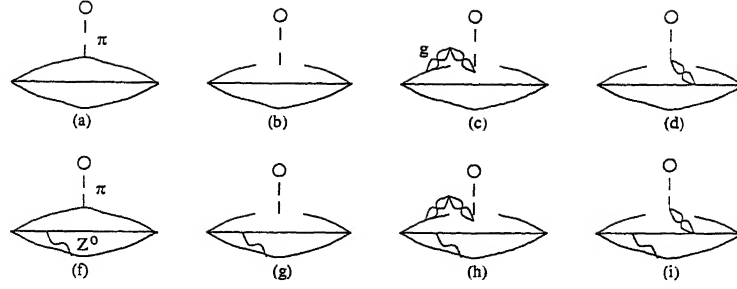
#### 3.3.2 External Pion Field for $g_{\pi N}$

The external field method was also used for determining the  $\pi - N$  strong and weak coupling constants, with the nucleon correlator in an external pion field:

$$\Pi_2^\pi(p) = i \int d^4x e^{ix \cdot p} \langle 0 | T[\eta_p(x) \bar{\eta}_p(0)] | 0 \rangle_\pi. \quad (72)$$

The weak parity-violating (PV) pion-nucleon coupling constant,  $f_{\pi N}$  is of particular interest because of its sensitivity to the neutral currents contribution of weak nonleptonic processes at low energies[76]. The theoretical prediction of  $f_{\pi N}$  is an important and challenging problem. The most accurate PV experiments have only shown[77, 76] that the upper limit for the magnitude of this coupling constant is 3-5 times smaller than the “best value” predicted by DDH[78] on the basis of a quark model and somewhat smaller than that in a similar calculation carried out more recently[79]. Since that time others have tried to estimate  $f_{\pi N}$  by means of chiral soliton models[80, 81] and QCD sum rules[82].

Recently the QCD sum rule external field method was used for the simultaneous calculation of both  $g_{\pi N}$  and  $f_{\pi N}$ , making use of an important observation illustrated in Fig. 12. In comparison of the lowest-dimensional diagrams for  $g_{\pi N}$  [Figs. 12a,b,c,d] and those for  $f_{\pi N}$  [Figs. 12e,f,g,h], note that all diagrams in which the weak gauge boson,  $Z^0$ , is exchanged between the quark coupled to the pion and one of the other quarks vanish! That is, the  $Z^0$  is exchanged only between spectator quarks. The weak  $W^\pm$  exchanges vanish as expected from known symmetries[78, 83]. This enables

Fig. 12 Lowest dimension diagrams for strong and weak  $\pi$ -N coupling

one to use the known value of  $g_{\pi N} \sim 13.4$  to estimate the most important unknown in the problem, the pion quark condensate,  $\chi^\pi$ . The result is that  $\chi^\pi a \simeq -1.88 \text{ GeV}^2$ , in agreement with the three-point estimate of Ref.[68]. The predicted value of the weak parity violating  $\pi - N$  coupling constant is  $f_{\pi N} = (3.04 \pm .01) \times 10^{-7}$ . A value close to this was obtained using a chiral quark model[72].

## 4 Form Factors and Reactions

To treat the elastic form factor of a hadron one must consider three-point functions. Since there is an extra variable, the momentum transfer  $q^2$  in comparison to the two-point correlators used to obtain masses, in the QCD sum rule approach one must deal with double dispersion relations,

$$F(Q_1^2, Q_2^2, Q^2) = \frac{1}{\pi^2} \int_0^\infty ds_1 \int_0^\infty ds_2 \frac{\rho(s_1, s_2, Q^2)}{(s_1 + Q_1^2)(s_2 + Q_2^2)}, \quad (73)$$

where  $\rho(s_1, s_2, Q^2)$  is the spectral function and the variables used are  $(Q_1^2, Q_2^2) = -(p_1^2, p_2^2)$ , the external momenta, and  $Q^2 = -q^2$ . For the phenomenological side of the sum rule, the elastic form factor is obtained by placing the external momenta on shell, so that  $s_1 = s_2 = M^2$ . For the study of nonperturbative as well as perturbative QCD one wishes to study  $F(Q^2)$  for all values  $0 \rightarrow \infty$ , and the challenge is to treat the low and medium values of  $Q^2$ .

### 4.1 Form Factors and Nonlocal Condensates

One approach is to treating the low and medium momentum transfer form factors is to introduce nonlocal condensates[84], as was done in a study of the pion form factor[85]. The nonlocal condensate, defined in Eq.(65), replaces the OPE for the nonperturbative quark propagator, as shown in Eq.(10). This enables one in principle to treat low  $Q^2$ . In effect the condensates carry momentum, as has been discussed in several publications[86, 87, 88].

The phenomenological side of  $F(Q_1^2, Q_2^2, Q^2)$  includes the pion elastic form factor by a term

$$\rho_\pi(m_\pi^2, m_\pi^2, Q^2) = \pi^2 f_\pi^2 F_\pi(Q^2) \delta(s_1 - m_\pi^2) \delta(s_2 - m_\pi^2), \quad (74)$$

which enables one to obtain sum rule expressions for  $F_\pi(Q^2)$ . The study in Ref.[85] shows that the soft part of the form factor (i.e., the triangle diagram dominates at about  $Q^2 \simeq 2 \text{ GeV}^2$ , and must be considered at least until  $Q^2$  is about  $10 \text{ GeV}^2$ . The program needed to carry out a detailed calculation is discussed.

#### 4.1.1 Pion Wave Function

The sum rule methods that we have discussed above start with a local composite field operator for each hadron, the current of the hadron. For example, for a  $\Gamma$  meson the current is  $\eta^\Gamma(x) =$

$\bar{q}(x)\Gamma q(x)$ , with  $\langle 0|\eta^\Gamma|H\rangle = \lambda^\Gamma v^\Gamma$ , where  $v^\Gamma$  is a spinor form and  $\lambda^\Gamma$  is a structure function depending on the wave function at the origin. In this form there is no direct information about the spacial or momentum form of the wave function. In Ref.[89] it was pointed out that by considering the gauge-invariant form

$$\langle 0|\bar{q}(x)\Gamma \exp[i g \int_0^x d\sigma A_\mu(\sigma)]q(0)|\Gamma(p)\rangle \quad (75)$$

one can study the wave function  $\phi_\Gamma(p)$ . For the pion the axial operator  $\Gamma = \gamma_\mu \gamma_5$  is used, and in Ref[89] the  $\pi^+$  was studied, so that the low-twist  $\pi^+$  wave function can be extracted from the matrix element

$$\langle 0|\bar{d}(x)\gamma_\mu\gamma_5 u(x)|\pi^+\rangle \quad (76)$$

By expanding in the gauge field,  $A_\mu$ , expressions for the moments of the wave functions

$$\langle \xi^n \rangle = \int_{-1}^1 d\xi \xi^n \psi(\xi)_\mu^2, \quad (77)$$

where  $\mu$  is the renormalization point, were obtained. In Ref.[90] the sum rule expressions were changed to make use of the nonlocal condensates. The moments that were obtained are very different from Ref.[89] and are similar to those obtained using the asymptotic wave function. This again suggests that for medium distances/medium momentum transfers nonperturbative effects must be considered. We return to the pion form factor and wave function in the next section

## 4.2 Light-Cone Sum Rules

It has long been known that for the relativistic study of composite states a light-cone representation of the field theory (or of Hamiltonian quantum mechanics) has many advantages[91, 92]. A Lorentz boost can be made with kinematic operators in a light-cone representation, while in the standard space-time representation the interactions are involved in all boosts. The light-cone QCD sum rules approach is an expansion in powers of the deviation from the light cone, rather than the O.P.E. of the original QCD sum rules[5]. The method is reviewed in Ref.[93]. The early applications were in heavy-quark physics[94, 95, 96]. Here we review the applications to the pion form factor and pion wave function.

### 4.2.1 Pion Form Factor and Light-Cone Sum Rules

The starting point of the light-cone sum rule method for the pion elastic form factor is similar to what we have referred to as the pion-vacuum correlator for the pion coupling constant, Eq(70), in that instead of starting with a vacuum matrix element one starts with a matrix element between the vacuum and one-pion state. In obtaining  $g_{\pi N}$  the pion is at zero momentum[75], while for the form factor one must consider finite  $q^2$ . This method allows one to treat the form factor as a two-point function rather than as a three-point function. The starting point[97] for the  $\pi^+$  form factor is

$$F_{\mu\nu}(p, q) = i \int d^4x e^{iqx} \langle 0|J_\mu^5(0)J_\nu^{em}(x)|\pi^+\rangle, \quad (78)$$

where  $J_\mu^5 = \bar{d}\gamma_\mu\gamma_5 u$ , the axial vector current, and  $J^{em}$  is the electromagnetic current. The pion form factor,  $F_\pi(Q^2)$  is obtained from the pole in the dispersion relation for the phenomenological side of  $F_{\mu\nu}$

$$F_{\mu\nu}^{pole} = 2i \frac{f_\pi F_\pi(q^2)}{m_\pi^2 - (p-q)^2} (p-q)_\mu p_\nu. \quad (79)$$

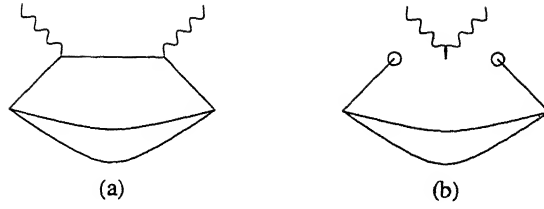


Fig. 13 D.I.S. (a) high-Q quark distributions, (b) sea-quark distributions

In Ref.[97] the form is expressed in terms of the pion wave function of Ref.[89] to leading order. The results depend on the model of the pion wave function, but the authors are in general agreement with Ref.[85] that soft physics must be considered for the pion form factor up to at least 10 GeV<sup>2</sup>.

#### 4.2.2 Pion Wave and Light-Cone Sum Rules

The light-cone sum rule method has also been used for a study of the pion wave function[98]. Starting with the expression for the low-twist pion wave function, Eq.(76), as in Ref.[89],

$$\langle 0 | \bar{d}(x) \gamma_\mu \gamma_5 u(x) | \pi^+(p) \rangle = i f_\pi p_\mu \int_0^1 du e^{-iu(px)} \phi_\pi(u) + \text{twist four} + O(x^2), \quad (80)$$

where  $\phi_\pi$  is the twist-two wave function. Going onto the light cone and treating the  $\gamma_+$  component, the left-hand side is expressed in terms of a model light-cone gaussian wave function, allowing one to extract  $\phi_\pi$  and the twist four light-cone wave function. The results of Ref.[98] are that the wave function, including the twist four part is similar to the asymptotic wave function, which is different from the original conclusion of Ref.[89],

#### 4.3 Deep Inelastic Scattering; Quark, Sea-Quark Distributions

Deep inelastic scattering (DIS) on nucleons is given by the lowest-twist mechanism, the “hand-bag diagram” illustrated in Fig. 13a. For electromagnetic DIS is treated by a four-point hadronic tensor

$$T_{\mu\nu}(p, q) = i \int d^4x d^4y d^4z e^{iqx} e^{ip(y-z)} \langle 0 | T[\eta(y) J^{em}(x) J^{em}(0) \bar{\eta}(z)] | 0 \rangle. \quad (81)$$

Such a four-point study of DIS to obtain quark distribution functions was carried out in the QCD sum rule framework for lepton nucleon DIS[88] in a calculation of valence quark distributions.

DIS can also be treated with a three point function, which is less complicated. This was done in Ref.[69]. The starting point is the twist-2 spin-averaged quark distribution,  $q(x, \mu^2)$ [99]

$$q(x, \mu^2) = \frac{1}{2\pi} \int d\sigma \langle p | \bar{q}(0) \gamma \cdot n \exp[ig \int_0^\sigma d\eta n \cdot A(\eta)] q(\sigma) | p \rangle_{\mu^2}. \quad (82)$$

Through an OPE one can show that

$$q(x, \mu^2) = \sum_{i=1}^{\infty} \frac{(-1)^{(i-1)}}{(i-1)!} \delta^{(i-1)}(x) \langle O_q^i \rangle_{\mu^2}, \quad (83)$$

with

$$O_q^i = \bar{q} \gamma \cdot n (i D \cdot n)^{(i-1)} q, \quad (84)$$

where  $D_\mu = \partial_\mu - igA_\mu$  is the covariant derivative.

Using the QCD sum rule method consider the three point correlator

$$T_q^i(p) = - \int d^4x d^4y e^{ip(x-y)} \langle 0 | T[\eta(x) O_q^i(0) \bar{\eta}(y)] | 0 \rangle. \quad (85)$$

The double pole term in the phenomenological dispersion relation is

$$T_q^{i,pole}(p) = \lambda_p^2 \frac{\not{p} + M_N}{(p^2 - M_N^2)^2} \langle O_q^i \rangle. \quad (86)$$

Using the typical sum rule methods, with the light-cone integrals discussed in Ref.[69], one can estimate  $\langle O_q^i \rangle$  and from this the quark distribution functions.

It was observed in Ref.[69] that the DIS on the quark condensate itself, shown in Fig. 13 b gives an estimate of the sea-quark distributions. By using nonlocal quark condensates defined in Eq.65 with  $g(y^2)$  of monopole form

$$g(y^2) = \frac{1}{(1 + \lambda^2 y^2/8)}, \quad (87)$$

one finds a sea-quark distribution in reasonable agreement with the empirical value[100]. Subsequently it was shown[68] that the monopole form does not have satisfactory analytic properties and a similar fit was made with a dipole form (Eq.66).

#### 4.3.1 Sea-Quark Anisotropy

The NMC/CERN experiments[101] gave evidence for the violation of the Gottfried sum rule and the recent Fermilab/E866[102, 103] Drell-Yan measurements show that for small Bjorken  $x$  the ratio  $\bar{d}(x)/\bar{u}(x)$ , down to up sea quark distributions, is considerably larger than 1.0, while both perturbative[104] and nonperturbative QCD calculations[69] find  $\bar{d}(x) \simeq \bar{u}(x)$ . In order to solve this problem there have been many theoretical calculations of DIS using the concept of a meson cloud based on the one-pion interaction of nucleons[105] (see Ref.[106] for a review). It is obvious that if one only considers the  $\pi^+$ -neutron system as an additional component of the proton, that there are more  $\bar{d}$  than  $\bar{u}$  sea quarks, but the  $\pi^- \Delta$  components tend to cancel this effect. A related model is the chiral quark model which has been used for flavor and spin properties of nucleons[107]. Within the QCD sum rule method a model correlator with a meson cloud has been proposed[108] and used for the study of nucleon magnetic dipole moments, but has not yet been applied to DIS.

## 5 Dyson-Schwinger, B-S Formalism and Sum Rules

The Dyson-Schwinger [D-S] for a fermion propagator[109] is an integral equation for the propagator in terms of the field with which the fermion interacts and the fermion-field vertex, all of which are dressed. For QCD, if we write the dressed (perturbative plus nonperturbative) inverse quark propagator as

$$S_q(p)^{-1} = i\not{p} + m_q + \Sigma(p), \quad (88)$$

then the D-S equation is

$$\Sigma(p) = \int \frac{d^4q}{(2\pi)^4} g_s^2 D_{\mu\nu}^{ab}(p-q) \gamma_\mu \frac{\lambda_c^a}{2} S_q(q) \Gamma_\nu^b(q,p) \quad (89)$$

with  $D_{\mu\nu}^{ab}(q)$  the dressed gluon propagator,  $\lambda_c^a$  the color  $SU(3)$  matrix and  $\Gamma_\nu^b$  the dressed gluon-quark vertex. D-S really consists of a coupled set of complicated integral equations involving the

quark and gluon propagators and vertex. This formalism has been widely used for the study of nonperturbative QCD in recent years[110, 111, 112]. The calculations have used models for  $\Gamma_\nu^b$ . This enables one to use known properties of the quark propagator, such as the condensates and the space-time structure of the nonlocal condensate, to constrain the Gluon propagator. For the study or the condensates often the “rainbow” approximation, with  $\Gamma_\nu^b(q, p) \Rightarrow \gamma_\nu \frac{\lambda^b}{2}$ . Working in a Feynman-like gauge, the gluon propagator has the form

$$D_{\mu\nu}^{ab}(q) = \delta^{ab} \delta_{\mu\nu} D(q), \quad (90)$$

With the  $D(q)$  a function chosen to give both long-distance confinement and ultra-violet behavior of the gluon propagator. An example is given below in the subsection on D-S and instantons quark propagators. Rewriting Eq.(88) as

$$S_q(p)^{-1} = i \not{p} A(p^2) + B(p^2) \quad (91)$$

The D-S equation can be written as a couples set of integral equations

$$\begin{aligned} [A(p^2) - 1]p^2 &= \frac{8}{3} g_s^2 \int \frac{d^4 q}{(2\pi)^4} D(p-q) \frac{A(q^2)}{q^2 A^2(q^2) + B^2(q^2)} p \cdot q \\ B(p^2) &= \frac{16}{3} g_s^2 \int \frac{d^4 q}{(2\pi)^4} D(p-q) \frac{B(q^2)}{q^2 A^2(q^2) + B^2(q^2)}. \end{aligned} \quad (92)$$

From the solutions for  $A(q^2)$  and  $B(q^2)$  one can obtain the quark condensate[113], the mixed condensate[114] and the quark condensate nonlocality[115] defined in Eq.(65):

$$\begin{aligned} \langle \bar{q}(0)q(0) \rangle &= -\frac{3}{4\pi^2} \int ds s \frac{B(s)}{sA^2(s) + B^2(s)} \\ \langle 0 | : \bar{q}(0)g\sigma \cdot G(0)q(0) : | 0 \rangle &= \frac{9}{4\pi^2} \int ds s \left[ s \frac{B(s)(2 - A(s))}{sA^2(s) + B^2(s)} + \right. \\ &\quad \left. \frac{81B(s)[2sA(s)(A(s) - 1) + B^2(s)]}{16(sA^2(s) + B^2(s))} \right] \\ g(y^2) &= (-)\frac{3}{4\pi^2} \int_0^\infty ds s \frac{B(s)}{sA^2(s) + B^2(s)} \left[ 2 \frac{J_1(\sqrt{sx^2})}{\sqrt{sx^2}} \right] \end{aligned} \quad (93)$$

With suitable choices of the function  $D(q^2)$  one can fit the condensates and the nonlocal condensate. This provides important constraints on the nonperturbative gluon propagator.

From Eqs.(92,93) one can see how the results of QCD sum rules can provide information for detailed model calculations of hadronic structure. The fit to the condensate information constrains the nonperturbative part of the gluon propagator, enabling one to carry out improved Bethe-Salpeter models of the hadrons. From this one can attempt much more detailed predictions of hadronic form factors, transition amplitudes, and so forth, than can be done with the QCD sum rules themselves. We anticipate that the productive interplay between the QCD sum rule and Bethe-Salpeter/Dyson-Schwinger studies will enable theorists to learn more about QCD from medium-energy experiments.

## 6 Instantons and QCD Sum Rules

A theory of the non-perturbative quark propagator is obtained by using the instanton solutions to the gluon field  $A_\mu^a$  and the solutions for a quark propagating in such an instanton medium[116]. In Refs. [5, 6] the instanton solutions for the gluon field were discussed in relation to the gluon condensate of the QCD sum rules; and the use of the quark propagator in the instanton field in the application of QCD sum rule method for hadronic properties was considered in the early days of the method[117, 118]. In these references very detailed discussions of the use of t'Hooft's

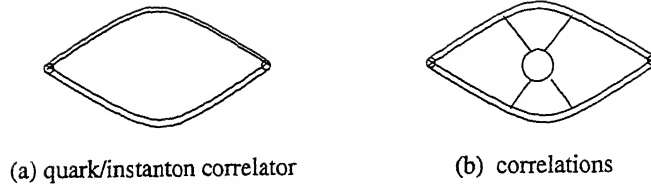


Fig. 14 Meson correlator with quark propagator in instanton medium

zero mode solutions were shown to be inconsistent with known applicability of a dilute instanton gas model, and it was concluded that interactions modifying the instantons must be included for meaningful hadronic physics. In Ref.[117] the idea of an instanton liquid model was discussed, but as pointed out there even in classical physics liquids are very complicated. Consequently, except for the possibility of large numerical lattice calculations, one must rely on models for attempting to include instanton effects into realistic hadronic and nuclear calculations. We discuss some work on microscopic QCD models below.

If one includes only the zero-mode contributions, the space-time form of quark propagator in the instanton medium is[116, 117]

$$S_q(x, y) = S_q^{PT} + \frac{\psi_0(x)\psi_0(y)}{m_{eff}}$$

$$\psi_0(x) = \frac{\rho \tilde{x}}{\pi x(x^2 + \rho^2)^{-3/2}}(1 - \gamma_5)\chi, \quad (94)$$

where  $\rho$  the instanton size,  $\chi^*\chi = 1$  and the instanton is assumed to be at the origin. The effective mass,  $m_{eff}$ , corresponds to the constituent quark mass of quark models. With this picture the QCD sum rule processes for a meson correlator would correspond to the diagrams shown in Fig. 14. Comparing Figs. 4 and 14, Fig. 14a is the correlator corresponding to Fig. 4a with the perturbative quark propagator replaced by the quark propagator in the instanton medium. Fig. 14b corresponds to correlations. If instantons provide all of the nonperturbative QCD effects the gluon condensate processes, Fig. 4b, would included in the processes of Fig. 14. We shall return to this below.

A self-consistent model of the quark propagator including modifications of the instanton medium by the light quarks has been developed[119], and consistency with the known value of the quark condensate,  $\langle \bar{q}q \rangle$ , and the constituent quark mass (about 345 MeV, consistent with quark models) is obtained. Meson correlators were treated using this model[120] and included correlations shown schematically in Fig. 14b. An improved model for the quark propagator in the instanton vacuum that goes beyond the approximation of Eq.(94) was developed[121] and an explicit form for  $S_q$  was found. With the notation used above that  $S_q(p)^{-1} = iA(p^2)\not{p} + B(p^2)$ , the instanton solution obtained for the inverse quark propagator is

$$A_I(p) - 1.0 = 0.0$$

$$B_I(p) = Kp^2 f^2\left(\frac{5}{6}p\right)$$

$$f(p) = \frac{2}{p} - (3I_0(p) + I_2(p)) \times K_1(p), \quad (95)$$

where  $K = 0.29 \text{ GeV}^{-1}$ ,  $p$  is in units of GeV and  $I_i, K_i$  are standard Bessel-type functions. This model predicts that the effective constituent quark mass at low momentum  $B_I(0) \simeq 420 \text{ MeV}$  with a  $1/p^6$  falloff. We return to this in a test of self-consistency using the D-S formalism.

## 6.1 Instanton Quark Propagator and QCD Sum Rules

Recently models of quark propagators in the instanton medium have been used in QCD sum rules. In a study of the nucleon correlator using the standard current of Eq.(32) but with a quark propagator of the form of Eq.(94) with  $m_{eff}$  taken as

$$m_{eff} = -\frac{2}{3}\pi^2\rho^2 <\bar{q}q>, \quad (96)$$

the two QCD sum rules for the coefficients of  $\bar{p}$  and  $I$ , Eqs.(33,34), were derived[122]. Although the instanton contributions to the sum rules were much smaller than the OPE contributions, the authors found a marked improvement in the stability of the sum rules. This is quite important as only with stable sum rules can one be assured of a good solution. In a related calculation[123] of the magnetic dipole moments of nucleons it was shown that the stability of the sum rules was improved, and that a third sum rule considered unreliable in the original effective field calculation of the moments[16] became stable.

## 6.2 D-S Study of Instanton Quark Propagator

As discussed in the previous section, the D-S formalism has been used to provide useful constraints on the nonperturbative gluon propagator from known properties of the quark condensates: the values of local and nonlocal quark condensates. Here we turn this around. By this formalism, given a form of the  $D(q^2)$  function, Eq.(90), with the confining long-distance as well as ultra-violet short distance form of the gluon propagator, and the gluon-quark vertex,  $\Gamma_\nu^b(q, p)$ , one can test the self-consistency of a model dressed quark propagator. The form of the  $D(q)$  function used in the D-S calculations[113, 114, 115] is

$$g_s^2 D(s) = 3\pi^2 \frac{X^2}{\Delta^2} e^{-\frac{s}{\Delta^2}} + c_u \frac{4\pi^2 d}{s \ln(\frac{s}{\Lambda^2} + e)}, \quad (97)$$

with the parameter  $c_u = (1.0, 0.0)$  to (include, neglect) the perturbative ultra-violet behavior. The strength parameter  $X$  and range parameter  $\Delta$  are determined by solving the coupled Dyson-Schwinger equations, Eq.(92), and fitting  $f_\pi$ , the pion decay constant, and the quark condensate through Eq.(93). For the Feynman gauge with  $c_u = 0.0$  these parameters are [115]  $X = 1.4 \text{ GeV}$  and  $\Delta = 2.0 \times 10^{-3} \text{ GeV}^2$ .

In a recent study using this D-S formalism[124] the instanton quark propagator derived in Ref.[121] and given in Eq.(95) was tested for self consistency using forms for  $D(q)$  that fit the condensates, such as the one just mentioned. The constant  $K$  in Eq.(95) was modified so that  $B_I(0) \simeq 313 \text{ MeV}$ , consistent with constituent quark models. The results of this calculation are that although the quark condensate can sometimes be self-consistently reproduced, the mixed condensate is in serious error. The mixed condensate has been estimated[125] from the solution of Ref.[121] and found to be about a factor of two too large. In Ref[124] the mixed condensate was found to be a factor of five too large.

These results suggest that the present models for the quark propagator in the instanton medium do not include all of the nonperturbative QCD needed for hadronic properties. It is known that present instanton models are not confining. This suggests that the current instanton quark propagator models can account for the short-distance perturbative and the mid-range nonperturbative QCD physics, but for the long-range confining nonperturbative effects gluonic condensates must still be included. This is consistent with the original observations[6] of the difficulty of including instanton effects in microscopic QCD calculations of hadronic properties.



## 7 Hadrons in Nuclear Matter

A main motivation for the study of hadrons on nuclear matter is the  $(\sigma, \omega)$  model [126, 127] which has been widely used in nuclear physics. It has long been known from the two-nucleon data that the N-N force has a short-range repulsion and a mid-range attraction (in addition to the long-range pion exchange force). The  $(\sigma, \omega)$  model extends this general idea to complex nuclei, with the effective interaction picture of the  $\sigma$  a scalar meson giving attraction and the  $\omega$  meson a chargeless vector meson being mainly repulsive. By picking parameters one can fit the binding energy of nuclear matter. In a mean field approximation the nucleon is treated as a Dirac particle an effective potential, and the effective mass of the nucleon in nuclear matter,  $M^*$ , is found to be reduced by 20-40% in various models.

The QCD sum rule method is a natural formalism for attempting to understand the physics of the effective nucleon mass in terms of QCD. Among other things there is a possibility of addressing the partial restoration of chiral symmetry at finite nuclear density, and the introduction of the concept of vacuum condensates at finite density is a natural approach, since the quark condensate is the parameter for chiral symmetry breaking. The pioneering work in using QCD sum rules at finite density [128] discussed the main ideas as well as the complications involved in such a program. Since then there have been a great deal of work by a number of authors on baryons and mesons in nuclear matter at finite density.

### 7.1 Nucleons in Nuclear Matter

The QCD sum rule treatment of a proton in nuclear matter have started with the nucleon current  $\eta$ , of Eq.(32), and a correlator defined with the vacuum state of Eq.(13) replaced by the nuclear state,  $|A\rangle$ ,

$$\Pi(q, P) = \frac{1}{\pi} \int d^4x e^{ix \cdot q} \langle A | T[\eta(x) \bar{\eta}(0)] | A \rangle, \quad (98)$$

with  $P$  the momentum of the entire nucleus. In addition to the work of Ref.[128], extensive studies have been carried out by the authors of Refs.[129] and [130]. For the vacuum correlator of the nucleon recall that there are two covariants,  $\tilde{q}$  and  $I$ , and therefore two sum rules. For the nucleon in the nucleus there is a third covariant,  $\tilde{P}$ , so that the in-medium correlator can be expressed as

$$\Pi(q, P) = \Pi^1(q, \nu) \tilde{q} + \Pi^2(q, \nu) I + \Pi^3(q, \nu) \tilde{P}, \quad (99)$$

and therefore there are three sum rules, which is of great help in finding solutions. It is also important to note that the in the medium correlator can be written as

$$\Pi(q, P) = \Pi(q, P)^{2q} + \Pi(q, P)^{4q} + \Pi(q, P)^{6q}, \quad (100)$$

with the superscripts representing two-quark, four-quark and six-quark matrix elements. For example,

$$\Pi(q, P)^{2q} = 2\epsilon_{abc}\epsilon_{a'b'c'} \langle A | \gamma^\mu \gamma_5 S_d^{cc'}(x) \gamma_5 \gamma^\nu Tr[\gamma_m u S_u^{aa'}(x) \gamma_\nu C S_u^{bb'}(x) T(x) C] | A \rangle. \quad (101)$$

The new dynamics involved with the microscopic QCD calculation of the correlator is that the quarks and glue propagate in nuclear matter, so that one can think of these particles in effective optical potentials as depicted in Fig. 15. All of the processes shown in Fig. 15 except 15(h) are two-quark terms, needed for the evaluation of  $\Pi(q, P)^{2q}$ . The ellipses indicate the new physics involved with propagation in the medium. Fig. 15b has a perturbative quark in the medium, Fig. 15c has an in-medium quark condensate, Figs.15e has an in-medium gluon condensate, and Figs 15f,g have in-medium mixed condensate processes. The in-medium condensates are evaluated to

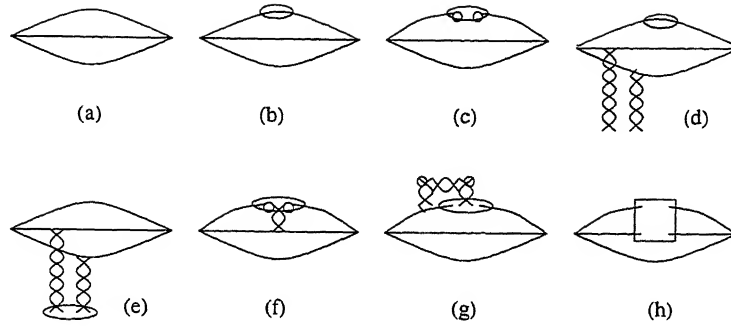


Fig. 15 Nucleon correlator in nuclear medium

first order in the nuclear density,  $\rho$ , and processes higher order in  $\rho$  are not included, recognizing that the nuclear medium is not very dense. The in-medium two-quark condensate has the form

$$\langle A | : \bar{q}^a q^b : | A \rangle = -\frac{\delta_{ab}}{12} [\langle \bar{q} q \rangle_\rho + \langle \bar{q} \not{P} q \rangle_\rho \not{P} / P^2]. \quad (102)$$

The important observation was made in the original work[128] that an estimate of the in-medium scalar density had been made to first order in  $\rho$  using PCAC[131]

$$\langle \bar{q} q \rangle_\rho = \langle \bar{q} q \rangle + \frac{\rho \sigma_N}{2m_q}, \quad (103)$$

where  $\sigma_N$  is the pion-nucleon sigma term. All of the calculations have used this result.

The great problem in carrying out this important program of research is the importance of four-quark condensates. A typical process is illustrated in Fig. 15h. A typical four-quark condensate term is

$$\Pi(q, P)^{4q} = 2\epsilon_{abc}\epsilon_{a'b'c'} \langle A | \gamma^\mu \gamma_5 S_d^{cc'}(x) \gamma_5 \gamma^\nu Tr[\gamma_\mu u^a(x) \bar{u}^{a'}(0) \gamma_\nu C u^{b'}(0) \bar{u}^b(x) C] | A \rangle \quad (104)$$

Generally some form of factorization is used to treat these in-medium matrix elements of four quark fields with various Dirac and SU(3) operators. In Ref.[128] and also in Refs.[129, 130] these difficulties have been discussed in some detail, but the solution has not been found. To illustrate how serious the problem is, one can find sum rule solutions for the mass of the nucleon increasing with density at low density or dropping to the values found in Walecka models depending on the model assumptions. For this reason we do not show results of calculations in this review. Some constraints on the four-quark condensates were found in a study of the  $\Delta(1232)$  in the medium[20], which will be briefly discussed in the next subsection.

## 7.2 Other Baryons in Nuclear Matter

The methods used for the study of the in-medium nucleon in nuclear matter[129] were applied to the treatment of the  $\Lambda(1115)$ . The current,  $\eta_\Lambda(x)$ , can be obtained from that of the nucleon, Eq.(32), by a transformation[7]. One main difference from the case of the nucleon is that the strange quark condensate in the medium must be used, which requires a modification of Eq.(103). The authors conclude once more that the method does not allow one to make predictions without solving the problem of the four-quark condensates, and once more different factorization schemes give very different results. One very interesting problem in  $\Lambda$  hypernuclear physics is the value of the spin-orbit interaction, and the authors point out that the QCD sum rule method could provide information.

The  $\Delta(1232)$  in nuclear matter has also been studied[132, 20]. In Ref[132] it was once more concluded that one could not make predictions without new knowledge of the many four-quark

condensates that are crucial for the in-medium QCD sum rule method. Ref[20] used quite a different approach: to attempt to put constraints on the in-medium four-quark condensates from experimental/theoretical knowledge of the mass and width of the delta in the nucleus. From the isobar doorway model fits to pion-nucleus elastic scattering and other reactions[133] one has learned that in finite nuclei the effective  $\Delta$  mass,  $M_\Delta^*$ , is almost the same as the free  $\Delta$  mass, with the best fit being about  $M_\Delta^* \simeq M_\Delta + 10$  MeV, while the width is broadened by about 10%. One must be careful in using these observations in that the  $\Delta$  tends to be formed in the nuclear surface at perhaps 1/2 nuclear density. Using the current  $\eta_\Delta(x)^\mu$ , given in Eq.(35), and defining the correlator  $\Pi^{\Delta*}_{\mu\nu}$  in analogy with Eq.(98), one can obtain three useful sum rules from the correlators

$$\Pi_1^{\Delta*} g_{\mu\nu} = Tr[\Pi^{\Delta*}_{\mu\nu}]/4, \quad (105)$$

and two other defined in Ref.[20]. The sum rule derived from the the correlator  $\Pi_1^{\Delta*}$  keeping processes as shown in Fig. 15 up to dimension seven is

$$\bar{\lambda}_{\Delta^*}^2 M_{\Delta^*} e^{-M_{\Delta^*}^2/M^2} = \frac{22}{4} a_\rho M^4 - \frac{3}{4} \bar{m}_0^2 a M^2 - \frac{7}{192} a_\rho b_\rho - \delta \Pi_1^{4q}, \quad (106)$$

with  $a_\rho = -(2\pi)^2 < \bar{q}q >_\rho$ ,  $b_\rho = < g^2 G^2 >_\rho$ ,  $\delta \Pi_1^{4q}$  a four quark condensate defined in Ref[20], and  $< 0|\eta_\Delta(x)^\mu|\Delta^* > = \bar{\lambda}_{\Delta^*} v_\mu / (2\pi)^2$ , with  $v_\mu$  a Rarita-Schwinger spinor. The other two sum rules are similar, but involve other in-medium condensates.

With these three sum rules it was possible to constrain some of the four-quark condensates. It was observed that some of the suggested factorizations were ruled out, and that with the constraint on one of the four-quark condensates found from the study of the in-medium  $\Delta$  the results for the in-medium nucleon would be significantly modified. The subject of baryons in nuclear matter is an important and interesting one, but is still an unsolved problem. In our section on mesons at finite T we return to the analysis of four-quark condensates.

### 7.3 Mesons in Nuclear Matter

Most of the theoretical and experimental research on mesons in nuclear matter, a very old subject in nuclear/particle physics, has been on vector mesons. Since the vector meson can couple directly to a photon, the decay of the in-medium  $\rho$  meson into lepton pairs provides a beautiful experimental test. There is a very large literature on the properties of the  $\rho$  meson in nuclei. Recently QCD sum rule calculations have been carried out[134, 135]. The problem of factorization of the four-quark condensates is still a major problem for this method. A recent review of light vector mesons in nuclear matter discusses the sum rule work as well as some of the other work in this area[136].

In the next section we take up the important area of hadrons in nuclei at finite temperature, and the recent work on the problem of treating four-quark condensates.

## 8 Mesons in Finite Temperature Matter

The main motivation for studying hadrons in nuclear matter is the possibility that in accelerator experiments it might be possible to create matter in the era of the early universe in which quarks and gluons are not confined, the quark-gluon plasma. Evidence from experiments at the AGS at BNL and the SPS at CERN that a hot, dense fireball is formed in relativistic heavy ion collisions[137], and signals that this is deconfined quark-gluon matter have been widely discussed[137]. In lattice gauge calculations (see, e.g., Ref.[138]) a phase transition from hadronic matter to deconfined matter occurs at a critical temperature of  $T = T_c \simeq 200$  MeV. See Ref.[118] for a review of this possible deconfining/chiral symmetry restoring phase transition.

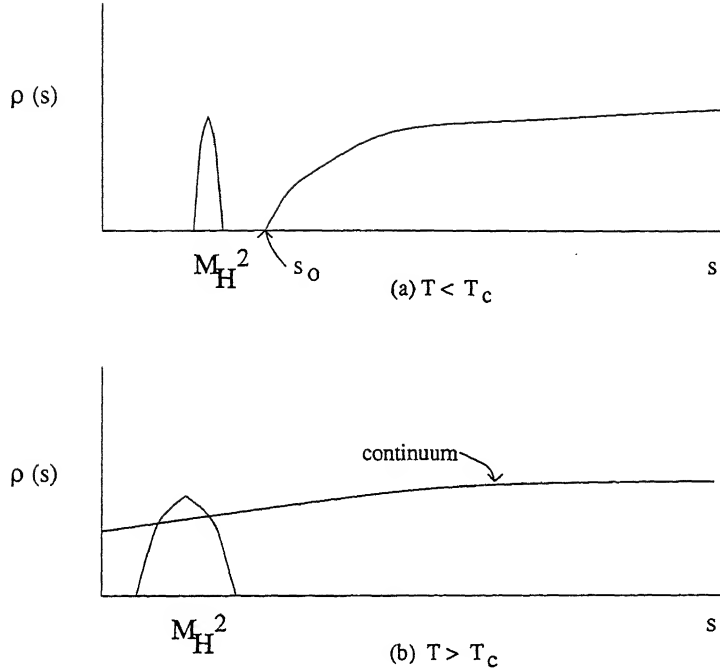


Fig. 16 Spectral functions at finite  $T$  for (a)  $T < T_c$  and (b)  $T > T_c$

As with the theory of mesons in nuclear matter, most of the research has been devoted to the study of vector mesons, since their decay into photons or lepton pairs make their experimental study most attractive. As we shall see, from the point of view of QCD sum rule theory, all of the mesons must be studied simultaneously in coupled equations. As has been emphasized in this review, one of the crucial aspects of the successful treatment of a correlator to extract information about the lowest meson (or baryon) state with a given set of quantum numbers by the QCD sum rule method is a satisfactory treatment of the continuum, the higher states with those quantum numbers. This is illustrated in Fig. 1, in which there is a clear separation in the spectral function between the lowest mass and the higher resonances and states, the continuum. The situation at finite  $T$  is illustrated in Fig. 16. For  $T < T_c$  as  $T$  begin to get near  $T_c$  the resonances become broader and the continuum spectral density drops toward the hadron mass being investigated. Therefore the criteria for stable solutions of the QCD sum rule equations might not be satisfied. For  $T > T_c$  the  $(q, \bar{q})$  states reach down to  $s=0$ , giving a spectral density illustrated in Fig. 16b. In that illustration it is assumed that mesons are beginning to form as  $T$  begins to approach  $T_c$  from the high- $T$  region, as found in a parton cascade models[139]. One can see that it is difficult to treat the region of temperature  $T \simeq T_c$  by the sum rule methods.

The pioneering work on using the QCD sum rule method[140] centered on the low- $T$  region. The starting point is the thermal-averaged correlator,  $\Pi^\Gamma(p, T)$ , which for a meson with a current  $J^\Gamma(x) = \bar{q}(x)\Gamma q(x)$  is

$$\Pi^\Gamma(p, T) = i \int d^4x e^{ix \cdot p} \theta(x_0) \langle \langle 0 | \bar{q}(x) \Gamma q(x) \bar{q}(0) \Gamma q(0) | 0 \rangle \rangle, \quad (107)$$

where  $p = (\omega, \mathbf{p})$  and where  $\langle \langle \dots \rangle \rangle$  stands for the Gibbs average. For finite- $T$  mesons the microscopic formulation differs from the  $T=0$  sum rule formulation in two ways: First, one replaces the perturbative quark loop process, shown in Fig. 4a, by a finite- $T$  quark loop using the Matsubara formalism in which for thermal equilibrium there is no time evolution and  $time \rightarrow i/T$ . In

Feynman diagrams the momentum  $p^\mu \rightarrow (i\omega_n, \mathbf{p})$ , where

$$\begin{aligned}\omega_n &\rightarrow 2\pi T n && \text{bosons (gluons), and} \\ \omega_n &\rightarrow 2\pi T(n + \frac{1}{2}) && \text{fermions (quarks),}\end{aligned}\quad (108)$$

and  $n$  runs over positive and negative integers (including zero). For example for the scalar case one finds the replacement for the first term in Eq.(27)

$$Q^2 \ln(Q^2) \rightarrow \int_0^{S_0} d\omega^2 \frac{1}{-p^2 + \omega^2} \tanh(\omega/4T). \quad (109)$$

Second, the condensates are evaluated in the medium. The gluon condensate does not seem to have a significant T-dependence. a calculation of the perturbative gluon loop[141] gives  $\langle G^2 \rangle_T^{P.T.} \simeq \frac{8\pi^2}{15} T^5$ . For  $T_c \simeq 200$  MeV one easily sees that  $\langle G^2 \rangle_{T_c}^{P.T.} \ll \langle G^2 \rangle / 10$ . Therefore for the region  $T < T_c$  the gluon condensate can be considered to be almost constant. This of course implies that the gluon condensate does not vanish as T becomes greater than  $T_c$ , which is most interesting.

Therefore the most significant T-dependent nonperturbative QCD effects for meson are the four-quark condensates:

$$\langle \hat{Q}_\Gamma \rangle_T \equiv \langle 0 | : \bar{q} \Gamma t^a q \bar{q} \Gamma t^a q : | 0 \rangle \rangle_T. \quad (110)$$

Bochkarev and Shaposhnikov used a model for the temperature dependence of the pertinent four-quark condensate, thus avoiding the factorization approximation, Eq.(30), according to which each four-quark condensate becomes trivially proportional to the square of the familiar quark condensate, assuming saturation by the vacuum. A number of other authors have also calculated the mass of the rho meson using QCD sum rules at finite T [142, 143], generally using factorization of the four-quark condensates. With the factorization the meson sum rule equations separate, which enables these authors to calculate the  $\rho$  mass without considering the other mesons. The results depend on assumptions about the continuum distribution and the T-dependence of the quark condensate, but with reasonable assumptions the  $\rho$  mass is found to drop with increasing T.

In recent work[145] a new formulation of QCD sum rules for mesons was developed by making use of the striking similarity of the correlator,  $\Pi^\Gamma(p, T)$  and the four quark condensates. Note that the spectral function of the correlator, Eq.(107), involves a retarded propagator, so that the relationship the the four-quark condensates, Eq.(110), requires a model of the meson intermediate states, as in Ref.[140]. Details are given in Ref.[145]. In this formulation one has sum rule expressions for both the meson masses and the four-quark condensates, which take the form

$$m_\Gamma^2(T) = H^\Gamma(T, m_\Gamma, S_0^\Gamma(T), M_B^\Gamma, \hat{Q}^\Gamma(T)), \quad (111)$$

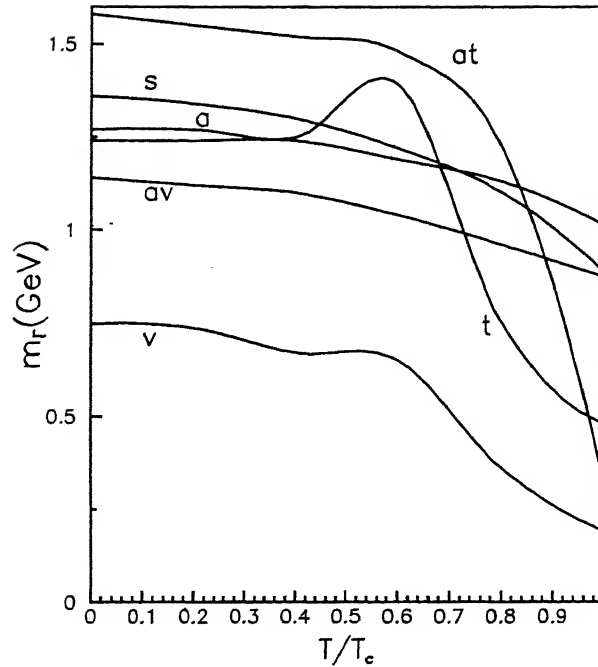
for the scalar, vector and axial vector meson masses, and

$$\langle \langle 0 | J^\Gamma J^\Gamma | 0 \rangle \rangle(T) = K^\Gamma(T, m(T), S_0(T), M_B^\Gamma), \quad (112)$$

for the five four-quark condensates, with  $M_B^\Gamma$  the Borel mass for each meson. In these equations there are five threshold functions,  $S_0(T)$ , which must be modeled, and  $\hat{Q}^\Gamma(T)$  are four-quark finite T condensates,

This formalism results in eight coupled equations. The models assumed for the threshold functions,  $S_0(T)$ , are described in Ref.[145]. Typical results for the meson masses are shown in Fig.17. The most striking results for masses are that the temperature dependences of the different mesons are quite different. Although the results depend on the spectral assumptions, the tensor and axial tensor mesons are found to be almost unstable as T approaches  $T_c$ , and this could provide an important experimental signal. The results for the four-quark condensates are shown in Fig.18. The functions  $R^\Gamma(T)$  in the figure are defined by

$$\langle \langle 0 | J^\Gamma J^\Gamma | 0 \rangle \rangle = R^\Gamma(T) \langle 0 | J^\Gamma J^\Gamma | 0 \rangle, \quad (113)$$

Fig. 17 Meson masses as a function of  $T$ 

where  $\langle 0 | J^\Gamma J^\Gamma | 0 \rangle$  are the  $T=0$  four-quark condensates for  $T=0$ . This figure, giving the ratio of  $R(T)$  for all the mesons relative to the scalar case, demonstrates the serious violation of factorization with increasing  $T$ . Vacuum saturation, the basis for factorization, implies that  $R^\Gamma(T)$  is the same for all mesons. We see that even at a modest temperature of  $T \approx T_c/2$  there is a considerable violation of factorization. One should note that at  $T=0$  there is evidence that the factorization approximation of the four-quark condensates is not accurate. In a study of the ratio of the isovector hadronic to the muon pair production in  $e^+e^-$  data [144] estimated a value of the vector four-quark condensate is larger than the factorized value. Using this value in Ref.[145] a satisfactory fit to the rho-meson mass was found. The QCD sum rule calculations that have been carried out for meson masses at finite  $T$  cannot be considered to be QCD predictions. They must make use of models to describe the physics as  $T$  approaches  $T_c$ . One hopes that with relativistic heavy ion experiments giving guidance the models will be improved and that in the future it might be possible to learn more about the chiral phase transition and QCD in the early universe.

## 9 The Pomeron

It has long been known that high energy elastic scattering and diffractive processes are dominated by the Pomeron trajectory. Since mesons are not on the Pomeron trajectory, the Pomeron must be a gluonic system. As we have seen in earlier sections, the QCD sum rule method is a valuable tool for studying gluonic hadrons, and I believe that the method will be very useful in understanding the structure of the pomeron and in making predictions for testing this nature. Although there is some recent work in this area, it is premature to include it in a review. It probably will be in reviews in this area in the future.

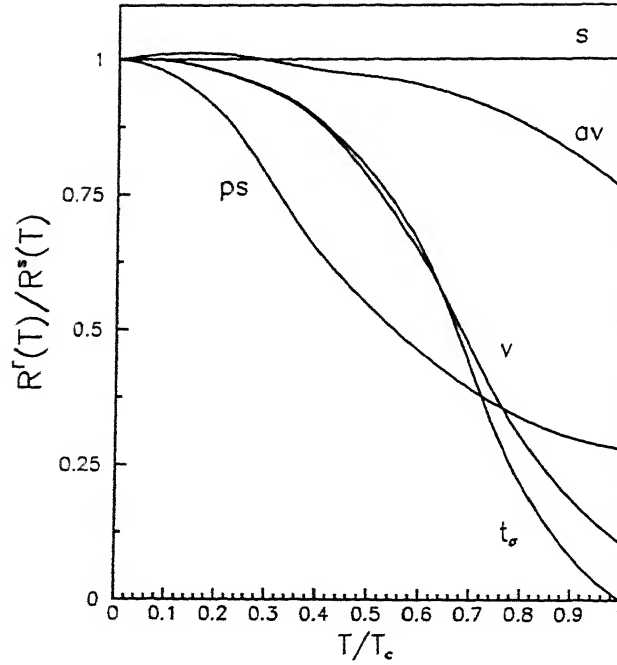


Fig. 18 Ratios of four-quark condensates as a function of  $T$

## 10 Outlook

The method of QCD sum rules will continue to be a valuable tool for the treatment of nonperturbative QCD in hadron spectroscopy, particularly for gluonic hadrons. There has been a long history of the study of glueballs and hybrids using the sum rule method. There is still more work to be done in this area. The mixing of mesons with glueballs and the nature of the possible light scalar glueball/sigma must be tested by experiments, and future experiments will give guidance for further theoretical work. A most important and interesting study will be that of the Pomeron. We expect that the use of instanton solutions for the quark propagators will be a valuable method for incorporating some of the developments in lattice gauge calculations in the treatment of nonperturbative QCD.

In all of these studies the most urgent research needed is for theory to provide reliable signals for gluonic systems, which requires more complicated correlators than the two-point correlators used for treating masses. There has been a great deal of progress in developing methods for treating coupling constants and other three point functions within the QCD sum rule method. The use of nonlocal condensates and the light-cone sum rule method are promising for theoretical studies of form factors and vertex functions over a wide range of momentum transfers. One promising area is in the interplay between QCD and electroweak theory, and the sum rule method could be particularly valuable in treating decays of heavy-light hadronic systems. We expect important developments in the near future in these areas.

Studies of hadrons in nuclear systems will be of increasing importance in the future. The QCD sum rule is most promising for several areas of nuclear astrophysics and cosmology. We anticipate a great deal of research on hadrons at finite temperature, with a valuable interplay between experiment and theory for the study of matter near the temperature of the chiral phase transition and of the nature of the quark-gluon plasma.

## 11 Acknowledgements

The author thanks Mikkel Johnson for many helpful discussions, and would like to acknowledge the hospitality of LANL during the time that much of this review was being prepared. The work was supported in part by the National Science Foundation grant PHY-9722143 and in part by the U.S. Department of Energy.

## References

- [1] R. Kronig, J. Opt. Soc. Am. **12**, 547 (1926); Physica **12**, 543 (1946).
- [2] H.A. Kramers, Atti con. int. fisici, Como, **2**, 545 (1927).
- [3] M. Gell-Mann, M.L. Goldberger and W.E. Thirring, Phys. Rev. **95**, 1612 (1954).
- [4] M. Gell-Mann, R.J. Oakes and B. Renner, Phys. Rev. **175**, 2195 (1968).
- [5] M.A. Shifman, A.I. Vainstein and V.I. Zakharov, Nucl. Phys. **B147**, 385, 448 (1979).
- [6] V. Novikov, M. Shifman, A. Vianstein, and V. Zakharov, Nucl. Phys. **B191**, 301 (1981).
- [7] L.J. Reinders, H.R. Rubinstein and S. Yazaki, Phys. Rep. **127**, 1 (1985).
- [8] T.M. Aliev and M.A. Shifman, Phys. Lett. **B112**, 401 (1982).
- [9] D.B. Leinweber, Annals Phys. **254**, 328 (1997), contains references to recent work.
- [10] E.V. Shuryak, Nucl. Phys. **B198**, 83 (1982)
- [11] T.M. Aliev and V.L. Eleskii, Sov. J. Nucl. Phys. **38**, 936 (1983).
- [12] N. Isgur and M.B. Wise, Phys. Lett. **B232**, 113 (1989).
- [13] M. Neubert, Phys. Rev. **D46**, 1076 (1992).
- [14] E. Bagan, p. Ball, V.M. Braun and H.G. Dosch, Phys. Lett. **B278**, 457 (1992).
- [15] L.S. Kisslinger and Z. Li, Nucl. Phys. **A570**, 167c (1994).
- [16] B.L. Ioffe, Nucl. Phys. **B188**, 317; **B191**, 591(E) (1981).
- [17] V.M. Belyaev and B.L. Ioffe, Sov. Phys. JETP **56**, 493 (1982).
- [18] D.B. Leinweber, Annals Phys. **198**, 203 (1990).
- [19] B.L. Ioffe, Z. Phys **C18**, 67 (1983).
- [20] M.B. Johnson and L.S. Kisslinger, Phys. Rev. **C52**, 1022 (1995).
- [21] D. Jido, N. Kodama and M.Oka, Phys. Rev. **D54**, 4532 (1996).
- [22] D. Jido, M.Oka, and A. Hosaka, Nucl. Phys. **A629**, c156 (1998).
- [23] V.A. Novikov, M.A. Shifman, A.I. Vainstein and V.I. Zakharov, Nucl. Phys. **B 165**, 67 (1980).
- [24] P. Pascual and R. Tarrah, Phys. Lett. **B 113** (1982) 495.
- [25] C.A. Dominguez and N. Paver, Zeit. Phys. **C 31**, 591 (1986).



- [26] J. Bordes, V. Gimenez and J.A. Penarrocha, Phys. Lett. **B 223**, 251 (1989).
- [27] E. Bagan and T.G. Steele, Phys. Lett. **243**, 413 (1990).
- [28] J. Liu and D. Liu, J. Phys. G:Nucl. Part. Phys. **19**, 373 (1993).
- [29] L.S. Kisslinger, J. Gardner and C. Vanderstraeten, Phys. Lett. **B410**, 1 (1997).
- [30] L.S. Kisslinger, Nucl. Phys. **A629**, 30c (1998).
- [31] S. Narison, Nucl. Phys. **B509**, 312 (1998).
- [32] S. Narison, N. Pak and N. Paver, Phys. Lett. **B147**, 162 (1984).
- [33] E. Bagan, A. Bramon and S. Narison, Phys. Lett. **B196**, 203 (1987).
- [34] G. Bali et al. (UKQCD), Phys. Lett. **B309**, 378 (1993); J. Sexton, A. Vaccarino and D. Weingarten, Phys. Rev. Lett. **75**, 4563 (1995).
- [35] Crystal Barrel Collaboration, Phys. Lett. **B 355**, 425 (1995).
- [36] C. Amsler and F.E. Close, Phys. Lett. **B 353**, 385 (1995); Phys. Rev. **D 53**, 295 (1996).
- [37] R.M. Baltrusaitis et al., Phys. Rev. Lett. **56**, 107 (1986).
- [38] BES collaboration, J.Z. Bai et al., Phys. Rev. Lett. **76**, 3502 (1996).
- [39] B.S. Zou and D.V. Bugg, Phys. Rev. **D50**, 591 (1994).
- [40] L.J. Reinders, S. Yazaki and H.R. Rubinstein, Nucl. Phys. **B 196**, 125 (1982).
- [41] I.I. Balitsky, D.I. Dyakonov and A.V. Yung, Phys. Lett. **B112**, 71 (1982).
- [42] J. Goverts, F. de Viron, D. Gusbin and J. Weyers, Nucl. Phys. **B248**, 1 (1984).
- [43] J.I. Latorre, S. Narison, P. Pascual and R. Tarrach, Phys. Lett. **B147**, 169 (1984).
- [44] D.R. Thompson et al. [BNL/E852 Collaboration], Phys. Rev. Lett. **79**, 1630 (1997).
- [45] F.E. Close and P.R. Page, Nucl. Phys. **B443**, 233 (1995).
- [46] T. Barnes *et al.*, Phys. Rev. **D52**, 5242 (1995).
- [47] J.I. Latorre, P. Pascual and S. Narison, Z. Phys. **C34**, 347 (1987).
- [48] A.P. Martynenko, Sov. J. Nucl. Phys. **54**, 488 (1991).
- [49] V.M. Braun, P. Gornicki, L. Mankiewicz and A. Schafer, Phys. Lett. **B 302**, 291 (1993).
- [50] L.S. Kisslinger and Z. Li, Phys. Rev. **D 51**, R5986 (1995).
- [51] L.S. Kisslinger and Z. Li, Phys. Lett. **B445**, 271 (1999).
- [52] K-C. Yang, W-Y.P. Hwang, E.M. Henley and L.S. Kisslinger, Phys. Rev. **D47**, 3001 (1993).
- [53] L.S. Kisslinger and Z. Li, Chinese J. Phys. **32**, 1213 (1994).
- [54] L.S. Kisslinger and Z. Li, Phys. Rev. Lett. **74**, 2168 (1995).
- [55] J. Gasser and H. Leutwyler, Phys. Rep. **87**, 77 (1982).

- [56] V.L. Eletsky and B.L. Ioffe, Phys. Rev. **D48**, 1441 (1993).
- [57] Particle Data Group, phys. Rev. **D45**, S1 (1992).
- [58] T. Goldman, K.R. Maltman and G.J. Stephenson Jr., Phys. Lett. **228**, 396 (1991).
- [59] L.S. Kisslinger, T. Goldman and Z. Li, Phys. Lett. **B416**, 263 (1998).
- [60] I.I. Balitsky and A. V. Yung, Phys.Lett. B **129**, 328 (1983).
- [61] B.L. Ioffe and A.V. Smilga, Nucl.Phys. B **232**, 109 (1984).
- [62] V. M. Belyaev and Ya. I. Kogan, Sov. J. Nucl. Phys. **40**, 659 (1984).
- [63] V. M. Belyaev and Ya. I. Kogan, JETP Lett. **37** (1983) 730; Phys. Lett. **136B** (1984) 273.
- [64] C. B. Chiu, J. Pasupathy and S. J. Wilson, Phys. Rev. D **32** (1985) 1786.
- [65] E. M. Henley, W-Y. P. Hwang and L. S. Kisslinger, Phys. Rev. D **46** (1992) 431.
- [66] H. He and X. Ji, Phys. Rev. **D54** (1996) 6897.
- [67] V. M. Belyaev and A. Oganesian, Phys. Lett. **B395** (1997) 307.
- [68] M.B. Johnson and L.S. Kisslinger, Phys. Rev. **D57**, 2847 (1998).
- [69] H. Jung and L.S. Kisslinger, Nucl Phys. **A586**, 682 (1995).
- [70] E. M. Henley, W-Y. P. Hwang and L. S. Kisslinger, Phys. Lett. **B367**, 21 (1996); **B440**, 449 (1998).
- [71] V. M. Belyaev and Ya. I. Kogan, Sov. J. Nucl. Phys. **40**, 659 (1984).
- [72] W-Y. P. Hwang, Z. Phys. **C75**, 701 (1997).
- [73] L.S. Kisslinger, hep-ph/9804320, Phys. Rev. C (1999).
- [74] L.J. Reinders, H.R. Rubinstein and S. Yazaki, Nucl. Phys. **B213**, 109 (1983); L.J. Reinders, Acta Phys. Polon. **B15**, 329 (1984).
- [75] H. Shiomi and T. Hatsuda, Nucl. Phys. **A594**, 294 (1995).
- [76] E.G. Adelberger and W.C. Haxton, Ann. Rev. Nucl. Part. Sci. **35**, 501 (1985); J. Lang *et al.*, Phys. Rev. **C34**, 1545 (1986).
- [77] C.A. Barnes *et al.* Phys. Rev. Lett. **40**, 840 (1978); H.C. Evans *et al.*, Phys. Rev. Lett. **55**, 791 (1985); Phys. Rev. **C35**, 1119 (1987); M. Bini *et al.*, Phys. Rev. Lett. **55**, 795, (1985); Phys. Rev. **C38**, 1195 (1988).
- [78] B. Desplanques, J.F. Donoghue, and B.R. Holstein, Ann. Phys.(NY) **124**,149 (1980).
- [79] V.M. Dubovik and S.V. Zenkin, Ann. Phys. (NY) **172**, 100 (1986).
- [80] N. Kaiser and U.G. Meissner, Nucl. Phys. **A499**, 699 (1989); **510**, 1648 (1990); U. Meissner, Mod. Phys. Lett. **A5**, 1703 (1990).
- [81] D.M. Kaplan and M.J. Savage, Nucl. Phys. **A556**, 653 (1993).
- [82] V.M. Khatsimovskii, Sov. J. Nucl. Phys. **42**,781 (1985).

- [83] E.M. Henley, N. Rev. Nucl. Sci. **19**, 367 (1969); Chinese J. Phys. **30**, 1 (1992).
- [84] S.V. Mikhailov and A.V. Radyushkin, Sov. J. Nucl. Phys. **49**, 494 (1989).
- [85] A.P. Bakulev and A.V. Radyushkin, Phys. Lett. **B271**, 223 (1991).
- [86] A.V. Kolesnichenko, Sov. J. Nucl. Phys. **39**, 968 (1984).
- [87] V.M. Belyaev and B.Y. Blok, Z. phys. **C30**, 279 (1986).
- [88] V.M. Belyaev and B.L. Ioffe, Nucl. Phys. **B310**, 548 (1988); **B313**, 647 (1989).
- [89] V.L. Chernyak and A.R. Zhitnitsky, Nucl. Phys. **B201**, 492 (1982); **B214**, 547(E) (1983).
- [90] S.V. Mikhailov and A.V. Radyushkin, Phys. Rev. **D45**, 1754 (1992).
- [91] P.A.M. Dirac, Rev. Mod. Phys. **21**, 392 (1949)
- [92] B.D. Keister and W.N. Polyzou, "Relativistic Hamiltonian Dynamics in Nuclear and Particle Physics", Adv. in Nucl. Phys. **20**, p. 225.
- [93] V.M. Braun, M. Beyer, T. Mannel and H. Schroder, Proceedings of the IVth International Workshop on Progress in Heavy Quark Physics, p.105.
- [94] I.I. Balitsky, V.M. Bruan and A.V. Kolesnichenko, Nucl. Phys. **B312**, 509 (1989)
- [95] V.M. Braun and I.E. Filyanov, Z. Phys. **C44**, 157 (1989).
- [96] V.L. Chernyak and I.R. Zhitnitskii, Nucl. Phys. **B345**, 137 (1990).
- [97] V. Braun and I. Halpern, Phys. Lett. **B328**, 457 (1994).
- [98] V.M. Belyaev and M.B. Johnson, Phys. Lett. **B423**, 379 (1998); Mod. Phys. Lett. **A13**, 2909 (1998).
- [99] R.L. Jaffe, Nucl. Phys. **B229**, 205 (1983).
- [100] D. Duke and J. Owens, phys. Rev. **D30**, 49 (1984).
- [101] P. Amaudruz et. al., Phys. Rev. Lett, **66**, 2712 (1991); M. Arneodo et. al., **D 50**, R1 (1994).
- [102] E. A. Hawker et. al., Phys. Rev. Lett. **80**, 3715 (1998).
- [103] J. C. Peng et. al., Phys. Rev. **D 58**, 92004 (1998)
- [104] D. A. Ross and C. T. Sachrajda, Nucl. Phys. **B149**, 497 (1979).
- [105] J. D. Sullivan, Phys. Rev. **D 5**, 1732 (1972).
- [106] J. Speth and A. W. Thomas, Adv. Nucl. Phys. **24**, 83 (1998).
- [107] T. P. Cheng and L.-F. Li, Phys. Rev. Lett. **74**, 2872 (1995); "Non-Perturbative QCD Spin Studies", hep-ph/9811279.
- [108] L.S. Kisslinger, hep-ph/9811497.
- [109] C. Itzykson and J-B. Zuber, "*Quantum Field Theory*" (McGraw-Hill Book Co., 1985).

- [110] C. D. Roberts and A. G. Williams, *Prog.Part.Nucl.Phys.* **33**, 477 (1994), and references therein.
- [111] P. Tandy, *Prog.Part.Nucl.Phys.***39**, 117 (1997) , and references therein.
- [112] A. Sharma and A.N. Mitra, hep-ph/9707503; A.N. Mitra, hep-ph/9906288; *Int. J. Mod. Phys. A***14**, 4589 (1999).
- [113] M.R. Frank and T. Meissner, *Phys.Rev. C* **53**, 2410 (1996).
- [114] T. Meissner, *Phys.Lett. B* **405**, 8 (1997).
- [115] L.S. Kisslinger and T. Meissner, *Phys. Rev. C***57**, 1528 (1998).
- [116] G. 't Hooft, *Phys. Rev. Lett.* **37**, 8 (1976); *Phys. Rev. D***14**, 3432 (1976).
- [117] E.V. Shuryak, *Nucl. Phys.* **B203**, 93; 116 (1982).
- [118] T. Schafer and E.V. Shuryak, *Rev. Mod. Phys.* **70**, 323 (1998).
- [119] D.I. D'yakonov and V.Yu. Petrov, *Sov. Phys. JETP* **62**, 204 (1985).
- [120] D.I. D'yakonov and V.Yu. Petrov, *Sov. Phys. JETP* **62**, 431 (1985).
- [121] P.V. Pobylitsa, *Phys. Lett.* **B226**, 387 (1989).
- [122] H. Forkel and M.J. Banerjee, *Phys. Rev. Lett.* **71**, 484 (1993).
- [123] M. Aw, M.K. Bannerjee and H. Forkel, hep-ph/9902458, to be published in *Phys. Lett. B* (1999).
- [124] L.S. Kisslinger, M. Aw, A. Harey and O. Linsuain, hep-ph/9906457.
- [125] M.V. Polyakov and C. Weiss, *Phys. Lett.* **B387** (1996) 841.
- [126] J.D. Walecka, *Ann. Phys.* **83**, 491 (1974).
- [127] B.D. Serot and J.D. Walecka, *Adv. in Nucl. Phys.* **16**, 1 (1986).
- [128] E.G. Drukarev and E.M. Levin, *ZhETF Lett.* **48**, 307 (1988); *Sov. Phys. JETP* **68**, 680 (1989); *Nucl. phys.* **A511**, 679 (1990), **A516**, 715(E) (1990); *Progress in Particle and Nuclear Physics*,**27**, 77 (1991).
- [129] T.D. Cohen, R.J. Furnstahl, and D.K. Griegel, *Phys. Rev. Lett.* **67**,961 (1991); X. Jin, M. Nielsen, T.D. Cohen, R.J. Furnstahl, and D.K. Griegel, *Phys. Rev. C***49**, 464 (1994), gives references to other publications of this group.
- [130] E.M. Henley and J. Pasupathy, *Nucl. Phys.* **A556**, 467 (1993).
- [131] T.P. Cheng, *Phys. Rev. D***13**, 2161 (1976).
- [132] X. Jin, *Phys. Rev. C***51**, 2260 (1995).
- [133] L.S. Kisslinger and W. Wang, *Phys. Rev. Lett.* **30**, 1071 (1973); *Ann. Phys. (N.Y.)* **99**, 374 (1976); A. Saharia, R. M. Woloshyn and L.S. Kisslinger, *Phys. Rev. C***23**, 2141 (1981).
- [134] T. Hatsuda and S.H. Lee, *Phys. Rev. C***46**, R34 (1992).
- [135] V.L. Eletsky and B.L. Ioffe, *Phys. Rev. Lett.* **78**, 1010 (1997).
- [136] T. Hatsuda, H. Shiomu and H. Kuwabara, *Prog. Theor. Phys.* **95**, 1009 (1996).
- [137] P. Braun-Munzinger and J. Stachel, *Nucl. Phys.* **A638** (1997) 3c.

- [138] T. Blum, *et al.*, Phys. Rev. **D51**, 5133 (1995).
- [139] K. Geiger and D.K. Srivastava, Phys. Rev. **C56**, 2718 (1997).
- [140] A. I. Bochkarev and M. E. Shaposhnikov, Nucl. Phys. **B268**, 220 (1986).
- [141] J.I. Kapusta, Nucl. Phys. **B148**, 461 (1979).
- [142] H. G. Dosch and S. Narison, Phys. Lett. **B203**, 155 (1988).
- [143] R. J. Furnstahl, T. Hatsuda and Su H. Lee, Phys. Rev. **D42**, 1744 (1990).
- [144] V. Giménez, J. Bordes and J. Peñarrocha, Nucl. Phys. **B357**, 3 (1991).
- [145] M.B. Johnson and L.S. Kisslinger, hep-ph/9908322

# 30. Light-Front Dynamics

V.A. Karmanov\*

Lebedev Physical Institute, Leninsky Prospekt  
53, 117924 Moscow, Russia

## Abstract

The wave function in relativity is defined, in four-dimensional space, on a space-like three-dimensional plane. The plane, most close to the time-like region, is the light-front plane  $ct + z = 0$ . Corresponding dynamical approach – *the light-front dynamics* – has considerable advantages. We describe, in a field-theoretical framework, the construction of light-front dynamics and illustrate it by some examples.

## 1 Introduction

A few centuries ago Galileo Galilei has discovered that the rectilinear motion is indistinguishable from the rest. Two observers, the laboratory observer and the moving one, carrying out the same experiments, obtain the same results. This discovery is deeply consistent with our intuition: the observer in an isolated laboratory does not interact with environment and, hence, he has no any way to learn about his motion.

At the beginning of this century the existence of the limiting velocity was established. This is the light speed  $c$ . Nothing can move faster. This discovery was also very consistent with our intuition. Indeed, if the limiting velocity would not exist, a very far part of the Universe could make an immediate influence to us. This seems unnatural.

According to the Galilei principle, the limiting speed should be the same in any moving system of reference. Otherwise, the observer would be able to notice his motion, measuring speed of light. However, this seems paradoxal from point of view of our everyday experience. Pursuing the light, we can accelerate our system almost until the light speed, but the light still runs away with the same speed  $c$ .

Einstein discovered, that the Galilei principle is reconciled with existence of the limiting velocity because of change of properties of space and time in a moving system relative to the rest one. For both observers the space-time in their own systems is the same, but for the observer from the rest system the space-time in the moving system looks different than his own one. In particular, when the speed  $v$  of the moving system approaches to  $c$ , the laboratory observer notices that the clock in this system delays from his one. In its turn, the observer in the moving system sees the similar effect: from his point of view the time in the rest system delays and almost stops when his speed approaches to  $c$ . Not only the clock, but any physical process observed from the moving system is stopped as well. To describe the physical phenomena, the laboratory observer can use, naturally, his own clocks and the space scales. However, on his choice, he can use the clocks and the space scales from the moving system. Two systems are equivalent, but two descriptions are different. The dilation of time can be used, in a theoretical laboratory, to make the "instant photo" of a fast, subnuclear physical process. "Stopping" the time, i.e., stopping the process, one obtains big advantage for study the most fast processes proceeding with the speed close to  $c$ . This dependence on the choice of the reference frame, is, in other words, the dependence on the choice of the space-time coordinates. In different coordinates the dynamical description of a system is different. We get in this way the different forms of dynamics.

---

\*Email: karmanov@sci.lebedev.ru

One of this form, the light-front dynamics (LFD), is very efficient tool to investigate the field theory and, in this framework, the relativistic composite systems (hadrons in the quark models, nuclei at relativistic relative nucleon momenta). In this article we will show, how LFD is constructed, explain its most principal properties, its relations to other approaches and give some applications. There are also a lot of phenomenological applications of LFD. They are beyond the scope of the present paper.

## 2 Forms of relativistic dynamics

In his famous article [1] Dirac analysed three forms of dynamics: the instant form, the point form and the front one.

From the group-theoretical point of view, the transformations of the system of reference including the translations, rotations and the Lorentz transformations are forming the Poincaré group. Under the infinitesimal transformation  $g$  of the coordinate system with the translation parameters  $a_\mu$  and with the four-dimensional rotation parameters  $\varepsilon^{\mu\nu}$ :

$$x^\mu \rightarrow x'^\mu = x^\mu + a^\mu + \varepsilon^{\mu\nu} x_\nu$$

the state vector  $\phi$  is transformed as follows:

$$\phi \rightarrow \phi' = U(g)\phi, \quad (1)$$

where

$$U(g) = 1 + iP_\mu a^\mu + \frac{i}{2} J_{\mu\nu} \varepsilon^{\mu\nu}. \quad (2)$$

Four translation generators  $P_\mu$  are the operators of the four-momentum. Six generators  $J_{\mu\nu}$  of the rotations and the Lorentz transformations are the operators of the four-dimensional angular momentum. The commutation relations between them have the form:

$$\begin{aligned} [P_\mu, P_\nu] &= 0, \\ \frac{1}{i}[P_\mu, J_{\kappa\rho}] &= g_{\mu\rho} P_\kappa - g_{\mu\kappa} P_\rho, \\ \frac{1}{i}[J_{\mu\nu}, J_{\rho\gamma}] &= g_{\mu\rho} J_{\nu\gamma} - g_{\nu\rho} J_{\mu\gamma} + g_{\nu\gamma} J_{\mu\rho} - g_{\mu\gamma} J_{\nu\rho}. \end{aligned} \quad (3)$$

The total angular momentum of the system is determined by the Pauli-Lubansky vector:

$$S_\mu = \frac{1}{2} \epsilon_{\mu\nu\rho\gamma} P^\nu J^{\rho\gamma}.$$

The state vector  $\phi^{J\lambda}(p)$  corresponding to a system with definite four-momentum  $p_\mu$ , mass  $M$ , total angular momentum  $J$  and its projection  $\lambda$  to the  $z$ -axis satisfies the following system of equations:

$$\begin{aligned} P_\mu \phi^{J\lambda}(p) &= p_\mu \phi^{J\lambda}(p), \\ P^2 \phi^{J\lambda}(p) &= M^2 \phi^{J\lambda}(p), \\ S^2 \phi^{J\lambda}(p) &= -M^2 J(J+1) \phi^{J\lambda}(p), \\ S_3 \phi^{J\lambda}(p) &= M \lambda \phi^{J\lambda}(p). \end{aligned} \quad (4)$$

A particular dynamical system is determined by the explicit form of these generators, i.e., by a particular solution of the commutation relations (3). If these generators are expressed in terms of the particle coordinates, we get a version of relativistic quantum mechanics with fixed number of particles. If the generators are expressed through the quantum fields, we obtain a form of the quantum field theory. As soon as the generators are known, the state vector is determined by eqs.(4). For an interacting system some Poincaré generators contain the interaction. Namely, the generators changing the position of the surface, where the state vector is defined, contain interaction. The generators, which do not change the position of the surface, don't contain interaction and coincide with the generators of free system. Using this property, one can classify the different forms of dynamics.

## 2.1 Instant form

The laboratory observer studies the physical processes in the four-dimensional space-time continuum described by the coordinates  $x = (t, \vec{r})$ . The three-dimensional space  $\vec{r}$  is a plane given by the equation  $t = \text{const}$ . The observer studies the evolution of his physical system from one plane  $t = \text{const}$  to other one. The wave function  $\psi(\vec{r}, t)$  of a quantum system, for a given  $t$ , is defined on this (three-dimensional) plane.

This description in four-dimensional space, from one equal-time plane to other one, corresponding to the different time instants  $t = \text{const}$ , is called the instant form of dynamics. In our everyday life we always use the instant form.

The time translations of the three-dimensional plane are determined by the Hamiltonian  $H = P_0$ . The interaction enters also into three operators of the Lorentz transformation  $J_{i0}$ ,  $i = 1, 2, 3$ . Indeed, two simultaneous events in one system of reference are not simultaneous ones in a moving system. Therefore, the Lorentz transformations don't leave the plane  $t = \text{const}$  invariant, they change the orientation of this plane relative to the time axis. This is the reason, why the corresponding generators contain the interaction.

Other six generators, the translations and rotations inside the three-dimensional space, namely,  $\vec{P}$  and  $\vec{J}_i = \epsilon_{ijk} J^{jk}$  coincide with the generators of the free system.

The instant form of dynamics is widely used for the relativistic generalizations of the quantum mechanics.

## 2.2 Point form

In principle, one can define the wave function not only on the plane, but on any space-like surface. Any two points of this surface can not be connected by the light signal and, hence, an event in one of these points cannot be a cause of the other one. A convenient choice is the surface of hyperboloid,  $t^2 - \vec{r}^2 = \text{const}$ . It is invariant under the Lorentz transformations. With the state vector defined on the family of these hyperboloids, we obtain the point form of dynamics.

In the point form the rotations and the Lorentz transformations don't change the hyperboloid  $t^2 - \vec{r}^2 = \text{const}$ . Therefore all the six generators  $J_{\mu\nu}$  don't contain the interaction. Whereas, the translations are much more complicated, and all the generators  $P_\mu$  contain the interaction. This means that the total momentum of a system is not the sum of the particle momenta. This complicates the situation, inspite of the simplification of the Lorentz boosts.

## 2.3 Front form

The observer moving with the velocity  $v$  along  $z$ -axis describes a physical process in his coordinates  $(t', x', y', z')$ , which are related to the laboratory ones by the Lorentz transformations:

$$\begin{aligned} z' &= \frac{z - vt}{\sqrt{1 - v^2/c^2}} \\ t' &= \frac{t + zv/c^2}{\sqrt{1 - v^2/c^2}} \\ x' &= x, \quad y' = y \end{aligned} \quad (5)$$

According to (5), the plane  $t' = \text{const}$  in moving system corresponds to  $t + zv/c^2 = \text{const}$  in the laboratory coordinates. The evolution is considered from one plane  $t + zv/c^2 = \text{const}$  to other one. Since the value of  $\text{const}$  is not yet specified, the factor  $1/\sqrt{1 - v^2/c^2}$  can be absorbed by it. For the "null plane" we put  $t' \propto t + zv/c^2 = 0$ . In the limiting case, when  $v \rightarrow c$ , we get the plane determined by the equation  $t' \propto z_+ = t + z/c = 0$ . The wave function is defined on this plane. This equation coincides with the equation for the light front  $z = -ct$ , moving along  $-z$ . This is the reason, why the description in these coordinates is called the front form of dynamics, or the light-front dynamics.

We emphasize that there are two equivalent points of view on LFD. On the one hand, we can study the system in the instant form, i.e., at  $t' = 0$ , but from point of view of the system of



reference moving with the limiting speed  $v \rightarrow c$ . This system of reference is called the "infinite momentum frame". One can equivalently describe the same system in the "normal", laboratory frame, but in the light-front coordinates  $(z_+, x, y, z_-)$ , here  $z_+ = t + z$  plays the role of the light-front "time",  $z_- = t - z$  is a coordinate in the light-front plane, and now we chose the unites with  $c = 1$ . The first approach is more convenient for intuition, the second one is more appropriate for technical developments. The both differ from the instant form,  $t = 0$ , in the laboratory system. The both should give, in principle, the same results, as the instant form, but, as we see, in more simple way.

From the group-theoretical point of view, in the front form of dynamics only three generators  $P_-, J_{1-}, J_{2-}$  do not leave the light-front plane invariant and contain the interaction. Other seven generators  $P_1, P_2, P_+, J_{12}, J_{-+}, J_{1+}$  and  $J_{2+}$  are the free ones.

Note also that, for a free particle, the relation between the energy and momentum  $p_0^2 = \vec{p}^2 + m^2$  can be rewritten in the light-front coordinates as:  $p_+ p_- - \vec{p}_\perp^2 = m^2$  (with  $\vec{p}_\perp = (p_1, p_2)$ ). So, the light-front energy  $p_-$  of a free particle is expressed through the momentum as:

$$p_- = \frac{\vec{p}_\perp^2 + m^2}{p_+}.$$

This expression does not contain any square root, in contrast to the instant form.

## 2.4 Why LFD?

The main difficulty of the quantum field theory is the very complicated structure of the state vector describing the particles and even the state without any particles – the vacuum state. The state vector is usually described as a superposition of the bare quanta, corresponding to the non-interacting fields. If we "switch off" the interaction between the fields, the number of particles is conserved. As soon as we take into account the interaction, the state vector is a superposition of the states with different numbers of particles.

If interaction is a weak, like in the case of the quantum electrodynamics, it does not change the state vector too much. Therefore, the "dressed" electron differs from the bare one only by small admixture of photon.

The situation is drastically different, when the interaction is strong. In this case, the structure of the real particle is extremely complicated. For example, the proton consists of three quarks, but these quarks are not the same quarks that appear in the initial Lagrangian of the Quantum Chromodynamics (QCD). They are so called the constituent quarks, which, in their turn, consist of the bare quarks and the gluons. The state vector of the proton is a huge superposition of the bare fields. It has not yet been calculated from the first principles of QCD.

One should emphasize that not only the proton state, but also the state without physical particles – the vacuum state, from point of view of the laboratory observer, is a complicated superposition of the bare particles, or, in other words, of fluctuations of the bare fields. At the same time, this description of emptiness in terms of the very complicated conglomerates of particles, seems unnatural. It would be much better to work in the approach, in which the vacuum is indeed nothing but emptiness. Simplifying the vacuum wave function, we simplify not only it, but also the wave function of the proton and of other particles, eliminating from them, like in the vacuum wave function, the fluctuations of fields. After that one can study the real, physical structure of particles.

*The vacuum is nothing but emptiness just in the light-front dynamics.* This is the principal advantage of this approach.

Qualitatively this can be understood from point of view of the uncertainty principle for energy and time. Consider the fluctuation creating three particles from vacuum. The fluctuation with the energy  $\Delta E = \varepsilon_{\vec{k}_1} + \varepsilon_{\vec{k}_2} + \varepsilon_{\vec{k}_3}$  may occur for the time  $\Delta t \approx \hbar/\Delta E$  (here  $\varepsilon_{\vec{k}} = \sqrt{\vec{k}^2 + m^2}$ ). In the infinite momentum frame the momenta  $\vec{k}_i$  and energies  $\varepsilon_{\vec{k}_i}$  of any particle increase,  $\Delta E$  tend to infinity. Therefore, the time of fluctuation  $\Delta t$  tends to zero. The contribution of this fluctuation to the vacuum wave function disappears.

This result is quite consistent with the mentioned above change of the space-time properties in the moving system. Due to the time dilation, all the physical processes are delayed, and the fluctuation has no time to occur. This means that in the thought experiment in the infinite momentum frame we study the particles prepared "far in advance", not spoiled by the vacuum fluctuations.

As already emphasized above, one can directly formulate the theory in the light-front variables, without taking any infinite momentum frame limit. This formulation includes the rules of the graph techniques, which allow to calculate the amplitudes. In principle, they could contain the vertices corresponding to vacuum fluctuations. We will see below that in LFD these vertices do not appear. This is the quantitative manifestation of the disappearance of the vacuum fluctuations. *In LFD, the bare vacuum state, i.e., the eigenstate of the free Hamiltonian, is also an eigenstate of full Hamiltonian, containing the interaction.* This property manifests itself in the formalism of LFD.

## 2.5 LFD and relativistic quantum mechanics

The dynamics of a nonrelativistic quantum system is determined by the Schrödinger equation with appropriate interaction Hamiltonian. Similar construction is developed for the relativistic quantum mechanical models. These models are based not on the field theory, but on a construction of relativistic phenomenological Hamiltonians in terms of the particle coordinates. The difference, in comparison to the nonrelativistic case, is in the fact that in the relativistic case the interaction enters in a few generators, so, we get a few "Hamiltonians". For example, in the front form, the "potential" is introduced in the generators  $P_-$ ,  $J_{1-}$ ,  $J_{2-}$ . It has to be introduced by a selfconsistent way, since the generators should satisfy the proper commutation relations of the Poincaré group. In this scheme one can fit the phenomenological potential, for example, between two nucleons, and then describe the properties of two-nucleon system: the deuteron wave function, the electromagnetic form factors, etc. The approach is also generalized to the three-body case. One can find the details in the review papers and books [2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. For the applications of the point form of dynamics to deep inelastic scattering see the paper [12].

Below in this article we concentrate on the field-theoretical approach in the framework of LFD. Many other details can be also found in the above review papers.

## 2.6 Explicitly covariant LFD

Together with big advantage of the simple vacuum structure, the light-front dynamics with the light-front plane  $z_+ = t + z = 0$  has a disadvantage: the coordinates  $x, y$  and  $z$  appear in a non-symmetric way. Because of that the theory loses the explicit relativistic and rotational covariance. For example, in the perturbation theory, the amplitude in a given order is determined by sum of a few time-ordered graphs which differ from each other by the relative time order of the interaction vertices. The sum of them is covariant, but any particular term in this sum is not covariant. So, we deal with the theory, which provides, in principle, the covariant final results, but not the intermediate ones. Because of approximations, the covariance of the final results can be also lost.

In spite of this inconvenience, LFD is applied in many papers to QCD, to the hadrons in quark models and to the relativistic nuclear physics. The applications to the light-front QCD and other references can be found, in particular, in [10, 13, 14]. Note that in the paper [15] it was shown that the constituent quark picture with logarithmic confinement naturally appears in weak coupling light-front QCD. The applications to the relativistic composite systems (hadrons and nuclei) and the corresponding references can be found in the above review papers. The rules of the graph techniques for the light-front quantum electrodynamics, alternative to the Feynman ones, were developed in [16, 17]. It has been demonstrated that the light-front QED reproduces the results obtained in the Feynman approach (such as anomalous magnetic moment of electron, etc.).

To avoid the inconvenience related to the absence of the covariance, the explicitly covariant version of LFD has been proposed [18] (see for review [11]). In this version the state vector is defined on the light-front plane of the general position, given by the equation  $\omega \cdot x = \omega_0 t - \vec{\omega} \cdot \vec{r} = 0$ ,

where  $\omega = (\omega_0, \vec{\omega})$  is a four-vector with  $\omega^2 = \omega_0^2 - \vec{\omega}^2 = 0$ . This is a generalization of the standard light-front approach. The latter corresponds to the particular value of  $\omega = (1, 0, 0, -1)$ .

The covariance means that, for example, any four-vector can be transformed from one system of reference to other one by a standard matrix, which depends on the kinematical parameters only, relating two system of reference. Therefore, this matrix is one and the same for all the four-vectors.

The absence of the explicit relativistic covariance in the standard version of LFD is related to the fact that the state vector depends dynamically on the orientation of the light-front plane. As mentioned above, the corresponding generators of these transformations contain the interaction. Rotating the system of reference, we rotate this plane. So, in the standard approach, with the light-front plane  $t + z = 0$ , there is no any universal kinematical transformation law for the light-front state vector.

In the explicitly covariant version of LFD the kinematical transformations of the system of reference are separated from the dynamical transformations of the light-front plane. So, all the transformations of the reference system are kinematical ones. This restores the explicit covariance. At the same time, the dependence of the state vector on the orientation of the light front remains to be dynamical. This orientation is determined by the direction of the four-vector  $\omega$ . The dependence of the state vector on the light-front orientation is now nothing but the dependence of the four-vector  $\omega$ . Therefore, the theory remains to be explicitly covariant.

In this scheme one can construct two sets of the Poincaré generators: (i) The generators responsible for transformations of the state vector under transformations of the reference system; they are kinematical and don't contain interaction. (ii) The generators responsible for transformations of the state vector under translations and rotations of the light-front plane; they are dynamical and contain interaction. The construction of these generators are given in Appendix. Group-theoretical aspects of the explicitly covariant LFD are clarified in the paper [19].

### 3 S-matrix

In the instant form, the S-matrix  $S(-\infty, t)$  gives the time evolution of the wave function, defined at  $t = -\infty$ , to the time  $t$ . The S-matrix  $S(-\infty, +\infty)$  gives the scattering amplitude. In LFD, this evolution takes place from one light-front plane to other one, in the direction of the light-front time.

As usual, the S-matrix is derived from the time-dependent Schrödinger equation in the "interaction representation":

$$i \frac{\partial \psi(t)}{\partial t} = H^{int}(t) \psi(t) \quad (6)$$

where

$$H^{int}(t) = \int H^{int}(\vec{x}, t) d^3x \quad (7)$$

is the interaction Hamiltonian,  $H^{int}(x) = H^{int}(\vec{x}, t)$  is the Hamiltonian density. We consider the example of the self-interacting scalar field:  $H^{int}(x) = -g\varphi^3(x)$ . In the interaction representation the field  $\varphi(x)$  is the free field:

$$\varphi(x) = \frac{1}{(2\pi)^{3/2}} \int \left[ a(\vec{k}) \exp(-ik \cdot x) + a^\dagger(\vec{k}) \exp(ik \cdot x) \right] \frac{d^3k}{\sqrt{2\varepsilon_k}}. \quad (8)$$

$a^\dagger, a$  are the creation and annihilation operators satisfying the commutation relation

$$[a(\vec{k}), a^\dagger(\vec{k}')] = (2\pi)^3 \delta^{(3)}(\vec{k} - \vec{k}').$$

The S-matrix is obtained as the formal solution of (6):

$$S = T \exp \left[ -i \int H^{int}(x) d^4x \right]. \quad (9)$$

The T-product orders the operators in the ordinary time  $t$ . The perturbation theory is obtained by decomposing (9) in series in the degrees of the coupling constant. One may put in correspondence, to any given term, a Feynman diagram and calculate the corresponding amplitude by the standard Feynman rules. In this way, the Feynman propagators appear as the average value, over the vacuum state, of the  $T$ -product:

$$G(x - x') = i \langle 0 | T(\varphi(x)\varphi(x')) | 0 \rangle.$$

Its Fourier transform is just the Feynman propagators:

$$\frac{i}{p^2 - m^2 + i0} = -i \int G(x) \exp(ipx) d^4x.$$

Another way to calculate the  $S$ -matrix is to develop the time-ordered perturbation theory. For this aim, following to [20, 21] (see for review [22]), one should replace in (9) the time-ordering operator  $T$  by the explicit time ordering. Namely, one can represent (9) as:

$$S = 1 + \sum_n \int (-i)^n H^{int}(x_1) \theta(t_1 - t_2) H^{int}(x_2) \dots \theta(t_{n-1} - t_n) H^{int}(x_n) d^4x_1 \dots d^4x_n. \quad (10)$$

In this way, the Feynman propagators are replaced by the average values of the product of the operators  $0|\varphi(x)\varphi(x')|0\rangle$ . There is no any  $T$ -product here, since it is taken into account by the theta-functions. In the momentum space, with

$$\tilde{\varphi}(k) = \frac{1}{(2\pi)^{5/2}} \int \varphi(x) \exp(-ik \cdot x) d^4x = [a(-\vec{k})\theta(-k_0) + a^\dagger(\vec{k})\theta(k_0)]\sqrt{2\varepsilon_k}\delta(k^2 - m^2)$$

this results in the contraction:

$$\underbrace{\tilde{\varphi}(k)\tilde{\varphi}(p)} = \tilde{\varphi}(k)\tilde{\varphi}(p) - :\tilde{\varphi}(k)\tilde{\varphi}(p): = \theta(p_0)\delta(p^2 - m^2)\delta^{(4)}(p + k). \quad (11)$$

We would like to emphasize that the propagator (11) contains the delta-function  $\delta(p^2 - m^2)$ , and therefore in the time-ordered graph techniques *all particles are always on their mass shells*. It is convenient to replace in the following  $\theta(p_0)$  in the propagator (11) by  $\theta(\omega \cdot p)$ . This is always possible, since  $p^2 = m^2 > 0$ .

This method results in the so called old fashioned perturbation theory. The amplitudes are represented by the time-ordered graphs. Instead of the Feynman propagators, they contain in the denominators the energies between the initial and intermediate states. The detailed derivation for arbitrary space-like plane is given in [20, 21, 22]. Namely, in the paper [20] the state vector is considered as evolving on the family of planes  $\lambda \cdot x = \sigma$ , where  $\lambda = (\lambda_0, \vec{\lambda})$ ,  $\lambda^2 = 1$ . The old fashioned perturbation theory is obtained from the graph techniques developed in [20, 21, 22] as a particular case at  $\lambda = (1, \vec{0})$ . The same method is applied to the case of ordering in the light-front time and gives the amplitudes in LFD. Below namely the latter case will be considered in detail. Here we illustrate in a simple example the difference between the Feynman and the usual time-ordered amplitudes. The amplitude for exchange by the particle in  $s$  channel can be represented in two different forms. The Feynman amplitude is:

$$M = \frac{g^2}{m^2 - (k + p)^2}$$

It corresponds to two terms in the old fashioned perturbation theory:

$$M = M_a + M_b = \frac{g^2}{2\varepsilon_{\vec{k}+\vec{p}} [\varepsilon_{\vec{k}+\vec{p}} - \varepsilon_{\vec{k}} - \varepsilon_{\vec{p}}]} + \frac{g^2}{2\varepsilon_{\vec{k}+\vec{p}} [\varepsilon_{\vec{k}+\vec{p}} + \varepsilon_{\vec{k}} + \varepsilon_{\vec{p}}]}. \quad (12)$$

Two items in (12) correspond to two time-ordered graphs, the second one arises from the vacuum fluctuation. It disappears in the infinite momentum frame (since  $\Delta E = \varepsilon_{\vec{k}} + \varepsilon_{\vec{p}} + \varepsilon_{\vec{k}+\vec{p}} \rightarrow \infty$ ) and in the light-front dynamics (see below).

Now consider the graph techniques, which is ordered in the light-front time. As mentioned, the LFD Hamiltonian is defined on the light-front plane  $\omega \cdot x = \sigma$ ,  $\sigma$  is the light-front time. Therefore, in the case of the scalar fields, the integral over  $d^3x$  in (7) is replaced by the integration over the light-front plane:

$$H^{int}(\sigma) = \int H^{int}(x) \delta(\omega \cdot x - \sigma) d^4x, \quad (13)$$

The S-matrix still has the form (10), but now the T-product orders the operators in the direction of  $\omega$ :

$$S = T_\omega \exp \left[ -i \int H_\omega^{int}(x) d^4x \right] \quad (14)$$

The expression (14) is then explicitly represented in terms of the light-front time  $\sigma = \omega \cdot x$ . Instead of (10) we get:

$$S = 1 + \sum_n \int (-i)^n H_\omega^{int}(x_1) \theta(\omega \cdot (x_1 - x_2)) H_\omega^{int}(x_2) \dots \theta(\omega \cdot (x_{n-1} - x_n)) H_\omega^{int}(x_n) \times d^4x_1 \dots d^4x_n. \quad (15)$$

The index  $\omega$  at  $H_\omega^{int}$  indicates that  $H^{int}$  and  $H_\omega^{int}$  may differ from each other in order to provide the equivalence between (9) and (14). The region where this can happen is a line on the light cone. Indeed, if  $(x_1 - x_2)^2 > 0$ , the signs of  $\omega \cdot (x_1 - x_2)$  and  $t_1 - t_2$  are the same and hence  $H_\omega^{int} = H^{int}$ . If  $(x_1 - x_2)^2 < 0$ , the operators commute:

$$[H^{int}(x_1), H^{int}(x_2)] = 0,$$

and their relative order has no significance. On the light cone, i.e. if  $(x_1 - x_2)^2 = 0$ ,  $\omega \cdot (x_1 - x_2)$  can be equal to zero while  $t_1 - t_2$  may be different from zero. If the integrand has no singularity at  $(x_1 - x_2)^2 = 0$ , this line does not contribute to the integral over the volume  $d^4x$ . However, if the integrand is singular, some care is needed. To eliminate the influence of this region on the S-matrix, we have introduced in (15) a new Hamiltonian  $H_\omega^{int}$ , such that expressions (9) and (15) be equal to each other. The form of  $H_\omega^{int}$ , which provides this equivalence, depends on the singularity of the commutator at  $(x_1 - x_2)^2 = 0$ . For the scalar fields, the singularity is weak enough, and the expressions (9) and (15) are the same, so that  $H_\omega^{int} = H^{int}$ . For fields with spins 1/2 and 1 or with derivative couplings, the equivalence is obtained with  $H_\omega^{int}$  differing from  $H^{int}$  by an additional contribution (counter term) leading to the contact terms in the propagators (or so called instantaneous interaction) [11].

Introducing the Fourier transform of the Hamiltonian:

$$\tilde{H}_\omega(p) = \int H_\omega^{int}(x) \exp(-ip \cdot x) d^4x, \quad (16)$$

and using the integral representation for the  $\theta$  function:

$$\theta(\omega \cdot (x_1 - x_2)) = \frac{1}{2\pi i} \int_{-\infty}^{+\infty} \frac{\exp(i\tau \omega \cdot (x_1 - x_2))}{\tau - i\epsilon} d\tau, \quad (17)$$

we can transform the expression (15) to the form:

$$S = 1 - i\tilde{H}_\omega(0) + \sum_{n \geq 2} (-i)^n \int \tilde{H}_\omega(-\omega\tau_1) \frac{d\tau_1}{2\pi i(\tau_1 - i\epsilon)} \tilde{H}_\omega(\omega\tau_1 - \omega\tau_2) \dots \frac{d\tau_{n-1}}{2\pi i(\tau_{n-1} - i\epsilon)} \tilde{H}_\omega(\omega\tau_{n-1}) \quad (18)$$

The  $\tau$  variable appears here as an auxiliary variable, as defined in eq.(17);  $\omega\tau$  has the dimension of a momentum.

### 3.1 Spin 0 system

Below we still restrict ourselves by the example of the simple interaction Hamiltonian of the form  $H = -g\varphi^3(x)$ . The covariant light-front graph technique arises when, as usual, one represents the expression (18) in normal form.

The four-vectors  $\omega\tau_j$  in (18) are associated with a fictitious particle – called *spurion* – and the factors  $1/(\tau_j - i\epsilon)$  are interpreted as the propagator of the spurions responsible for taking the intermediate states off the energy shell. This spurion should be interpreted as a convenient tool in order to take into account off-energy shell effects in the covariant formulation of LFD (in the absence of off-mass shell effects), and not as a physical particle. It is absent, by definition, in all asymptotic, on-energy shell states. We shall show below on simple examples how the spurion should be used in practical calculations.

The general invariant amplitude  $M_{nm}$  of a transition  $m \rightarrow n$  is related to the  $S$ -matrix by:

$$S_{nm} = 1 + \frac{i(2\pi)^4 \delta^{(4)}(\sum_{i=1}^m k_i - \sum_{i=1}^n k'_i)}{((2\pi)^3 2\varepsilon_{k'_1} \dots (2\pi)^3 2\varepsilon_{k'_n} (2\pi)^3 2\varepsilon_{k_1} \dots (2\pi)^3 2\varepsilon_{k_m})^{1/2}} M_{nm}, \quad (19)$$

where, e.g.,  $\varepsilon_{k_1} = \sqrt{m_1^2 + \vec{k}_1^2}$ . The cross-section of the process  $1 + 2 \rightarrow 3 + \dots + n$  is thus expressed as:

$$d\sigma = \frac{(2\pi)^4}{4j\varepsilon_{k_1}\varepsilon_{k_2}} |M|^2 \frac{d^3 k_3}{(2\pi)^3 2\varepsilon_{k_3}} \dots \frac{d^3 k_n}{(2\pi)^3 2\varepsilon_{k_n}} \delta^{(4)}(k_1 + k_2 - k_3 - \dots - k_n), \quad (20)$$

where  $j$  is the flux density of the incident particles:

$$j\varepsilon_{k_1}\varepsilon_{k_2} = \frac{1}{2}[s - (m_1 + m_2)^2]^{1/2}[s - (m_1 - m_2)^2]^{1/2}, \quad s = (k_1 + k_2)^2.$$

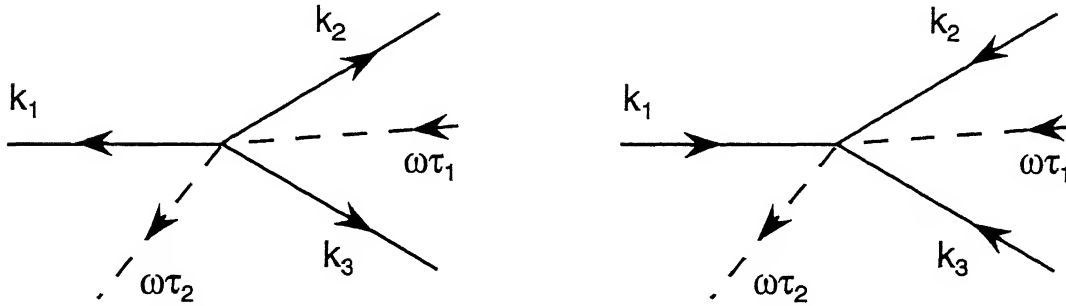


Fig.1. The vacuum vertices.

To find the matrix element  $M$  of order  $n$  one must proceed as follows [20, 21, 22, 18, 23]:

1. Arbitrary label by a number the vertices in the Feynman graph of order  $n$ . Orientate continuous lines (the lines of physical particles) in the direction from the smaller to the larger number. Initial particles are oriented as incoming into a graph, and final particles as outgoing. Connect by a directed dashed line (the spurion line) the vertices in the order of decreasing numbers. Diagrams in which there are vertices with all incoming or outgoing particle lines (vacuum vertices, as indicated in fig. 1) can be omitted. Associate with each continuous line a corresponding four-momentum, and with each  $j$ -th spurion line a four-momentum  $\omega\tau_j$ .
2. To each internal continuous line with four-momentum  $k$ , associate the propagator  $\theta(\omega \cdot k) \delta(k^2 - m^2)$ , and to each internal dashed line with four-momentum  $\omega\tau_j$  the factor  $1/(\tau_j - i\epsilon)$ .
3. Associate with each vertex the coupling constant  $g$ . All the four-momenta at the vertex, including the spurion momenta, satisfy the conservation law, i.e., the sum of incoming momenta is equal to the sum of outgoing momenta.

4. Integrate (with  $d^4k/(2\pi)^3$ ) over those four-momenta of the internal particles which remain unfixed after taking into account the conservation laws, and over all  $\tau_j$  for the spurion lines from  $-\infty$  to  $\infty$ .
5. Repeat the procedure described in 1-4 for all  $n!$  possible numberings of the vertices.

We omit here the factorial factors that arise from the identity of the particles and depend on the particular theory.

The important property of LFD – the disappearance of the vacuum fluctuations is just the disappearance of the vacuum vertices indicated in fig. 1. In this formalism they disappear for a trivial reason: it is impossible to satisfy the four-momentum conservation law for them. Indeed, the conservation law for the vertex of fig. 1 has the form  $k_1 + k_2 + k_3 = \omega(\tau_1 - \tau_2)$ . Since the four-momenta are on the mass shell:  $k_{1-3}^2 = m^2 > 0$ , so that the left-hand side is always strictly positive:  $(k_1 + k_2 + k_3)^2 \geq 8m^2$ , whereas the right-hand side is zero since  $\omega^2 = 0$ . However, it will be seen that the vacuum contributions that vanish in the light-front approach leave their track in a different way, making for the fields with spin the light-front interaction  $H_\omega(x)$  in eq.(10) different from the usual interaction  $H(x)$  in (9).

The case of the particles with non-zero spins is considered in [11]. In this case, the vacuum fluctuations disappear too, but some additional (contact) vertices appear, due to the difference between  $H^{int}$  and  $H_\omega^{int}$ . They are also taken into account by the rules of the graph techniques.

We emphasize that despite the presence of the four-vector  $\omega$  in eq.(18), the amplitudes calculated in this way are explicitly covariant. We just obtain the theory with separation of the kinematical dependence of amplitudes on the reference system and of the dynamical, but covariant dependence on the light-front orientation. The full  $S$ -matrix and any physical amplitudes do not depend on  $\omega$ , since eq.(18) gives the same  $S$ -matrix, as the initial one, given by eq.(10). However, off-shell amplitudes depend on  $\omega$  and off-shell light-front amplitudes don't coincide with the Feynman ones. We will see below, that the wave functions also depend on  $\omega$ .

The light-front diagrams can be interpreted as time-ordered graphs. As soon as the vertices are labelled by numbers, any deformation of a diagram changing the relative position of the vertex projections on the “time direction” from left to right does not change the topology of the diagram and the corresponding amplitude. Therefore it is often convenient to deform the diagram so that the vertices with successively increasing numbers are disposed from left to right. This just corresponds to time ordered graphs. In addition, this graph technique is three-dimensional one, i.e., the four-momenta of the particles, even in the intermediate states, are always, on the mass shells, all the integrations over the internal momenta are three-dimensional ones.

The light-front amplitudes can be also obtained from the graph techniques [20, 21, 22] with  $\lambda = (\lambda_0, \vec{\lambda})$ ,  $\lambda^2 = 1$  as follows. One should replace  $\lambda \rightarrow \lambda'/\delta$  with  $\lambda'^2 = \delta^2$  and take limit  $\delta \rightarrow 0$ . This just corresponds to the infinite momentum frame limit of the old-fashioned perturbation theory. The light-front amplitudes can be also obtained by direct transformation of a given Feynman amplitude [24, 25].

By a replacemet of variables [11] the covariant light-front amplitudes can be transformed to the form of the ordinary light-front diagrams corresponding to  $\omega = (1, 0, 0, -1)$ , given by the Weinberg rules [26].

### 3.2 Why time-ordered graphs?

Deriving both the Feynman graph techniques and the time-ordered one, we proceed from one and the same expression (9) for the  $S$ -matrix and therefore we obtain the same amplitude in a given order of the perturbation theory. The important difference between two approaches appears in describing the bound states, and, in general, the state vector. In the Feynman approach the bound states are described by the Bethe-Salpeter functions [27], which are defined as:

$$\Phi(x_1, x_2, p) = \langle 0 | T(\varphi(x_1)\varphi(x_2)) | p \rangle. \quad (21)$$

Here  $\varphi(x)$  is the Heisenberg operator. The Bethe-Salpeter function depends on two four-vectors  $x_{1,2}$ , they include two times  $t_{1,2}$ . In the momentum space the Bethe-Salpeter function looks as:

$\Phi = \Phi(l_1, l_2, p)$ . Their arguments  $l_{1,2}$  are off mass shell:  $l_1^2 \neq m^2$ ,  $l_2^2 \neq m^2$ . Though it satisfies the normalization condition, allowing to find the normalization coefficient, the Bethe-Salpeter function has no any probabilistic interpretation (see for review [28]).

The time-ordered approach describes the bound states by means of the Fock components. It allows to express the amplitudes in terms of the Fock componets of the state vector. The latters are the direct relativistic generalization of the non-relativistic wave functions. They depend on the on-mass-shell four-vectors and have the same probabilistic interpretation, as the non-relativistic wave functions. The kernel of the equation for the wave function can be calculated by the rules of the graph techniques. The time-ordered graphs give also the space-time picture of the process.

The transparant physical interpretation, clear nonrelativistic limit and also comparatively simple three-dimensional calculating formalism are the advantages of this approach.

The relation between the light-front wave function and the Bethe-Salpeter amplitude is given below in sect. 4.5.

### 3.3 Simple examples

#### 3.3.1 Exchange in $t$ -channel

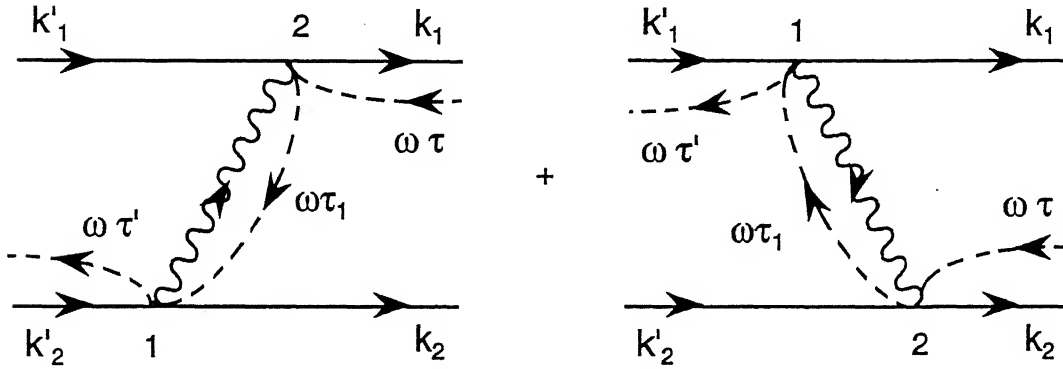


Fig.2. Exchange by a particle in  $t$ -channel.

Consider two time-ordered diagrams shown in fig. 2. They correspond to the exchange of a scalar particle of mass  $\mu$  between two scalar particles, in the  $t$  channel. These diagrams determine, in the ladder approximation, the kernel of the equation for the calculation of the light-front wave function. The external spurion lines indicate that the amplitude is off-energy shell. The term "off-energy shell", is borrowed from the old fashioned perturbation theory, where it means that for an amplitude which is an internal part of a bigger diagram, there is no conservation law for the energies of the incoming and outgoing particles (like in the intermediate states in the amplitudes (12)). For the light-front amplitudes shown in fig. 2, for  $\omega = (1, 0, 0, -1)$ , there is no conservation law for the minus-components of the particle momenta, i.e., for the "light-front" energies. This momentum nonconservation is just taken into account by spurion.

According to the light-front graph technique for spinless particles, the amplitude has the form:

$$\begin{aligned}
 \mathcal{K} &= g^2 \int \theta(\omega \cdot (k_1 - k'_1)) \delta((k_1 - k'_1 + \omega \tau_1 - \omega \tau)^2 - \mu^2) \frac{d\tau_1}{\tau_1 - i\epsilon} \\
 &\quad + g^2 \int \theta(\omega \cdot (k'_1 - k_1)) \delta((k'_1 - k_1 + \omega \tau_1 - \omega \tau')^2 - \mu^2) \frac{d\tau_1}{\tau_1 - i\epsilon} \\
 &= \frac{g^2 \theta(\omega \cdot (k_1 - k'_1))}{\mu^2 - (k_1 - k'_1)^2 + 2\tau \omega \cdot (k_1 - k'_1) - i\epsilon} \\
 &\quad + \frac{g^2 \theta(\omega \cdot (k'_1 - k_1))}{\mu^2 - (k'_1 - k_1)^2 + 2\tau' \omega \cdot (k'_1 - k_1) - i\epsilon}.
 \end{aligned} \tag{22}$$



The two items in (22) correspond to the two diagrams of fig. 2. They cannot be non-zero simultaneously. On the energy shell, i.e. for both  $\tau = \tau' = 0$ , the expression for the kernel is identical to the Feynman amplitude:

$$\mathcal{K}(\tau = \tau' = 0) = \frac{g^2}{\mu^2 - (k_1 - k'_1)^2 - i\epsilon}. \quad (23)$$

Note that the off-shell amplitude (22) depends on  $\omega$ .

On the energy shell, corresponding to  $\tau = \tau' = 0$ , the dependence of the amplitude on  $\omega$  disappears. In more complicated cases, when a Feynman diagram corresponds to the sum a few light-front diagrams (like in the case of the box diagrams considered in sect. 6 below), the amplitude for a particular light-front diagram may depend on  $\omega$  even on the energy shell. This dependence disappears in the sum of all amplitudes in a given order. In this case the singularities of different amplitudes, related to their dependence on  $\omega$ , cancel each other in the sum.

The dependence of the perturbative amplitude (22) on the light-front orientation (calculated exactly in the  $g^2$  order) indicates that the light-front wave function, being the off-shell object too, also depends inevitably on the light-front orientation (see sect. 4 below).

### 3.3.2 Self-energy contributions

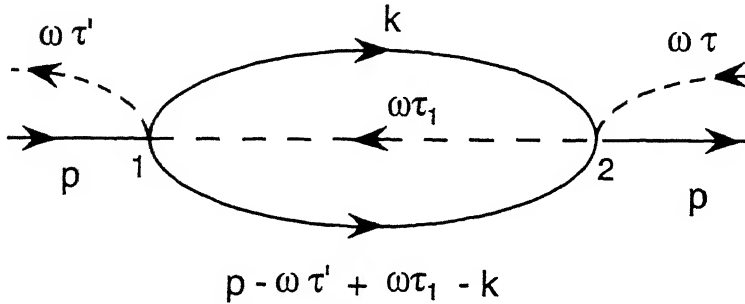


Fig.3. Self-energy loop.

Another simple example is the self-energy diagram shown in fig. 3. The corresponding amplitude (equal to the self-energy up to a factor) has the form:

$$\Sigma(p') = g^2 \int \theta(\omega \cdot k) \delta(k^2 - m^2) \theta(\omega \cdot (p' + \omega \tau_1 - k)) \delta((p' + \omega \tau_1 - k)^2 - m^2) \frac{d^4 k}{(2\pi)^3} \frac{d\tau_1}{\tau_1 - i\epsilon}, \quad (24)$$

with  $p' = p - \omega \tau'$ .

Let  $q = p' + \omega \tau_1$ . The integral over  $d^4 k$  is thus reduced to the well known calculation of the imaginary part of the Feynman amplitude, when all the propagators are replaced by the delta-functions:

$$\int \delta(k^2 - m^2) \delta((q - k)^2 - m^2) d^4 k = \frac{\pi}{2\sqrt{q^2}} \sqrt{q^2 - 4m^2}.$$

Inserted in (24), it gives:

$$\Sigma(p') = \frac{g^2}{16\pi^2} \int_{4m^2 - p'^2}^{\infty} \frac{\sqrt{p'^2 - 4m^2 + \tau_1}}{\sqrt{p'^2 + \tau_1}} \frac{d\tau_1}{\tau_1 - i\epsilon}. \quad (25)$$

The logarithmic divergence is at the upper limit of the integration over  $\tau_1$ . One can introduce the invariant cutoff in terms of  $\tau_1$ . In this way, after renormalization, the standard expression for the self-energy amplitude is obtained.

The finite value of  $\Sigma(p')$  for finite  $\tau_1$  is a particular manifestation of a general property of the light-front amplitudes. A peculiarity of the covariant light-front amplitudes is that they have no

any ultraviolet divergences for the finite values of all the spurion four-momenta. All the ultraviolet divergences in all the light-front diagrams appear after integrations over  $\tau_j$  in infinite limits [20]. Indeed, the energy-momentum conservation (including the spurion four-momentum) is valid in any vertex. Since all the four-momenta are on the corresponding mass shells, we have at each vertex a real physical process as far as the kinematics is concerned. For finite initial particle energies and for finite incoming spurion energy, the energies of the particles in the intermediate states are thus also finite. Hence, the integrations over the particle momenta for fixed spurion momenta are constrained by a kinematically allowed finite domain. It is the same reason that provides finite imaginary part of a Feynman diagram found by replacing the Feynman propagators  $\frac{1}{(k^2 - m^2 + i\epsilon)}$  by the delta-functions  $-i\pi\delta(k^2 - m^2)$ . In both cases the internal particle lines are associated with the delta-functions.

The only source of the ultraviolet divergences in the light-front amplitudes is the infinite intermediate spurion energies, i.e., infinite  $\tau_j$ . This is the reason why divergences may appear at the upper limit of integration over  $\tau_j$ . Since  $\tau_j$  are scalar quantities, one can introduce an invariant cutoff in terms of these variables. This way of regularizing the divergent diagrams is another advantage of the covariant formulation of LFD.

For the massless particles, the light-front amplitudes may have infrared divergences, like in the case of the Feynman diagrams.

Another peculiarity of LFD is the appearance of “zero modes”. For constituents of zero mass, for instance, the state vector may contain components with  $\omega \cdot k = 0$  for non-zero four-momentum  $k$ . In the standard approach, this corresponds to the finite light-front energy  $k_- = \vec{k}_\perp^2 / k_+$  for both  $k_+ = 0$  and  $\vec{k}_\perp^2 = 0$ . Zero modes can also appear in theories with spontaneously broken symmetry. They make the equivalence between LFD and the instant form of quantization in which nontrivial vacuum structures (condensates) appear [10, 29, 30, 31].

The detailed discussion of these important problems is beyond the scope of the present paper.

## 4 Light-front wave function

As already mentioned, the wave functions are the Fock components of the state vector defined on the light-front plane  $\omega \cdot x = 0$ . This means that they are coefficients in an expansion of the state vector  $|p\rangle$  with respect to the basis of free fields:

$$\begin{aligned} |p\rangle_\omega &\equiv \phi_\omega(p) \equiv (2\pi)^{3/2} \int \psi_2(k_1, k_2, p, \omega\tau) a^\dagger(\vec{k}_1) a^\dagger(\vec{k}_2) |0\rangle \\ &\times \delta^{(4)}(k_1 + k_2 - p - \omega\tau) 2(\omega \cdot p) d\tau \frac{d^3 k_1}{(2\pi)^{3/2} \sqrt{2\varepsilon_{k_1}}} \frac{d^3 k_2}{(2\pi)^{3/2} \sqrt{2\varepsilon_{k_2}}} + \dots \end{aligned} \quad (26)$$

The dots  $\dots$  include the higher Fock states. For simplicity, we omit the spin indices.

We emphasize in (26) the presence of the delta-function  $\delta^{(4)}(k_1 + k_2 - p - \omega\tau)$ . This gives the conservation law:

$$k_1 + k_2 = p + \omega\tau. \quad (27)$$

In the particular case where  $\omega = (1, 0, 0, -1)$ , the delta-function  $\delta^{(4)}(k_1 + k_2 - p - \omega\tau)$  gives the standard conservation laws for the  $(\perp, +)$ -components of the momenta, but does not constrain the minus-components.

From (26) one can see that the wave function depends on  $\omega\tau$ , i.e., on the orientation of the light front. This important property of any Fock component is very natural. As explained above, any off-energy shell amplitude depends on the light-front orientation (see eq.(22)). The bound state wave function is always an off-shell object ( $\tau \neq 0$ ). Therefore it also depends on the orientation of the light-front plane. This property is not a peculiarity of the covariant approach. At the same time, the description of the off-energy shell effects in terms of the external spurion lines allows to parametrize this dependence explicitly.

#### 4.1 The relativistic relative momentum

We will mainly concentrate on the two-body wave function. Generalization to the  $n$ -body case is straightforward and is given in [11].

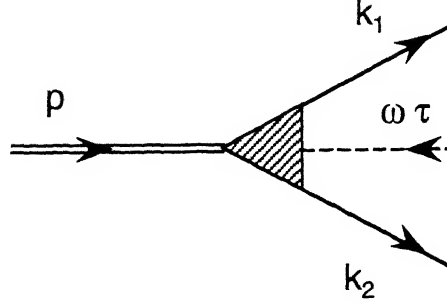


Fig.4. Graphical representation of the two-body wave function on the light front. The broken line corresponds to the spurion (see text).

Due to the conservation law (27), the light-front wave function can be shown graphically like a two-body scattering amplitude as indicated in fig. 4. The broken line corresponds to the fictitious spurion.

Due to this analogy, the decomposition of the wave function in independent spin structures and their parametrization is analogous to the expansion of a two-body amplitude in terms of invariant amplitudes. We will use this analogy below. We emphasize again that although we assign a momentum  $\omega\tau$  to the spurion, there is no any fictitious particle in the physical state vector. *The basis in eq.(26) contains the particle states only.*

The relativistic relative momentum  $\vec{k}$  has the same sense as the norelativistic one: it is the momentum of one of the particle in the c.m.-system where  $\vec{k}_1 + \vec{k}_2 = 0$ . Note that *due to the conservation law (27), the total momentum  $\vec{p} \neq 0$  of the system in this reference frame is not zero.* This definition of the relative momentum does not assume, however, that we restrict ourselves to this particular reference frame. In the arbitrary system of reference the relative momentum is constructed by the Lorentz transformation to the system moving with velocity

$$\vec{v} = \vec{P}/P_0, \quad \text{where } P = k_1 + k_2 = p + \omega\tau.$$

We get:

$$\vec{k} = L^{-1}(\mathcal{P})\vec{k}_1 = \vec{k}_1 - \frac{\vec{P}}{\mathcal{M}} \left[ k_{10} - \frac{\vec{k}_1 \cdot \vec{P}}{\mathcal{M} + P_0} \right], \quad (28)$$

$L^{-1}(\mathcal{P})$  is the Lorentz boost,  $\mathcal{M} = \sqrt{P^2}$ . Similarly we define the unit vector  $\vec{n}$  in the direction of  $\vec{\omega}$  in this system:

$$\vec{n} = L^{-1}(\mathcal{P})\vec{\omega}/|L^{-1}(\mathcal{P})\vec{\omega}| = \mathcal{M}L^{-1}(\mathcal{P})\vec{\omega}/\omega \cdot p. \quad (29)$$

From these definitions, it follows that under a rotation and a Lorentz transformation  $g$  of the four-vectors from which  $\vec{k}$  and  $\vec{n}$  are formed, the vectors  $\vec{k}$  and  $\vec{n}$  undergo only rotations:

$$\vec{k}' = R(g, \mathcal{P}) \vec{k}, \quad \vec{n}' = R(g, \mathcal{P}) \vec{n},$$

where  $R$  is the rotation operator:

$$R(g, p) = L^{-1}(gp)gL(p). \quad (30)$$

Therefore  $\vec{k}^2$  and  $\vec{n} \cdot \vec{k}$  are the rotation and the Lorentz invariants. For the wave function with zero angular momentum we thus obtain [18]:

$$\psi = \psi(\vec{k}, \vec{n}) \equiv \psi(\vec{k}^2, \vec{n} \cdot \vec{k}). \quad (31)$$

It is seen from (31) that the relativistic light-front wave function depends not only on the relative momentum  $\vec{k}$  but on another variable – the unit vector  $\vec{n}$ .

In the case of the states with non-zero angular momentum, the angular momentum is constructed by means of the spherical functions depending on the arguments  $\vec{k}$  and  $\vec{n}$ .

We introduce another set of variables in which the wave function can be parametrized, in analogy to the equal-time wave function in the infinite momentum frame. We define the variables:

$$x = \omega \cdot k_1 / \omega \cdot p, \quad R_1 = k_1 - xp, \quad (32)$$

and represent the spatial part of  $R$  as  $\vec{R} = \vec{R}_\parallel + \vec{R}_\perp$ , where  $\vec{R}_\parallel$  is parallel to  $\vec{\omega}$  and  $\vec{R}_\perp$  is orthogonal to  $\vec{\omega}$ . Since  $R \cdot \omega = R_0 \omega_0 - \vec{R}_\parallel \cdot \vec{\omega} = 0$  by definition of  $R$ , it follows that  $R_0 = |\vec{R}_\parallel|$ , and, hence,  $\vec{R}_\perp^2 = -R^2$  is invariant. Therefore,  $\vec{R}_\perp^2$  and  $x$  can be chosen as two the scalar arguments of the wave function:

$$\psi = \psi(\vec{R}_\perp^2, x). \quad (33)$$

Using the definitions of the variables  $\vec{R}_\perp^2$  and  $x$ , we can readily relate them to  $\vec{k}^2$  and  $\vec{n} \cdot \vec{k}$ :

$$\vec{R}_\perp^2 = \vec{k}^2 - (\vec{n} \cdot \vec{k})^2, \quad x = \frac{1}{2} \left( 1 - \frac{\vec{n} \cdot \vec{k}}{\varepsilon_k} \right). \quad (34)$$

The inverse relations are

$$\vec{k}^2 = \frac{\vec{R}_\perp^2 + m^2}{4x(1-x)} - m^2, \quad \vec{n} \cdot \vec{k} = \left[ \frac{\vec{R}_\perp^2 + m^2}{x(1-x)} \right]^{1/2} \left( \frac{1}{2} - x \right). \quad (35)$$

The variables introduced above can be easily generalized to the case of different masses and an arbitrary number of particles [23]. The corresponding variables  $\vec{q}_i, \vec{n}$  are still constructed according to eqs.(28), (29) and the variables  $\vec{R}_{i\perp}, x_i$  according to (32).

## 4.2 Normalization

The state vector is normalized as:

$${}_\omega \langle p', \lambda' | p, \lambda \rangle_\omega = 2p_0 \delta^{(3)}(\vec{p} - \vec{p}') \delta^{\lambda' \lambda}. \quad (36)$$

The Fock components are normalized so as to provide the condition (36). Substituting the state vector (26) in the left-hand side of eq.(36), we reproduce the right-hand side if  $\sum_n N_n^{\lambda' \lambda} \equiv \delta^{\lambda' \lambda}$ , where  $N_n^{\lambda' \lambda}$  is the contribution to the normalization integral from the  $n$ -body Fock component.

For the state with zero total angular momentum the normalization condition has the form:

$$\sum_n N_n = 1. \quad (37)$$

In this case, the two-body contribution to the normalization integral reads:

$$N_2 = \frac{1}{(2\pi)^3} \int \psi^2(\vec{k}, \vec{n}) \frac{d^3 k}{\varepsilon_k} = \frac{1}{(2\pi)^3} \int \psi^2(\vec{R}_\perp, x) \frac{d^2 R_\perp dx}{2x(1-x)}. \quad (38)$$

This normalization integral gives contribution only of the two-body wave function to the sum (37). The contribution of other sectors can be taken into account by the integral:

$$\frac{1}{(2\pi)^3} \int \frac{d^3 k}{\varepsilon_k} \frac{d^3 k'}{\varepsilon_{k'}} \psi^*(\vec{k}', \vec{n}) \left[ \varepsilon_k \delta(\vec{k} - \vec{k}') - \frac{4m^2}{(2\pi)^3} \frac{\partial V(\vec{k}', \vec{k}, \vec{n}, M^2)}{\partial M^2} \right] \psi(\vec{k}, \vec{n}) = 1, \quad (39)$$

where  $V(\vec{k}', \vec{k}, \vec{n}, M^2)$  is the kernel of the equation for the wave function. The second term accounts for the many-body contribution to the norm,  $\sum_{n>2} N_n$ .

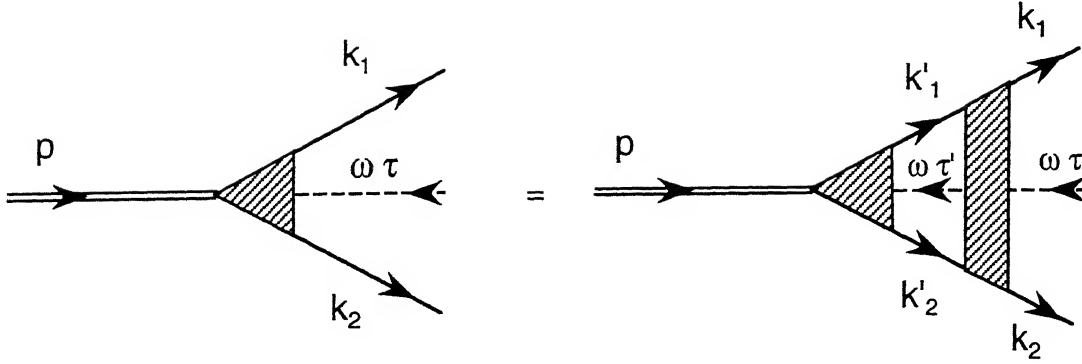


Fig.5. Equation for the two-body wave function.

### 4.3 Equation for the wave function

The equation for the wave function is obtained from the equation for the vertex part shown graphically in fig. 5.

It is the analogue, for a bound state, of the Lippmann-Schwinger equation for a scattering state. Let us first explain its derivation for the case of spinless particles. In accordance with the rules given in sect. 3.1, we associate with the diagram of fig. 5 the following analytical expression:

$$\begin{aligned} \Gamma(k_1, k_2, p, \omega\tau) &= \int \Gamma(k'_1, k'_2, p, \omega\tau') \theta(\omega \cdot k'_1) \delta(k'^2_1 - m^2) \theta(\omega \cdot k'_2) \delta(k'^2_2 - m^2) \\ &\times \delta^{(4)}(k'_1 + k'_2 - p - \omega\tau') d^4 k'_1 \mathcal{K}(k'_1, k'_2, \omega\tau'; k_1, k_2, \omega\tau) \frac{d\tau'}{\tau' - i\epsilon} \frac{d^4 k'_2}{(2\pi)^3}. \end{aligned} \quad (40)$$

Here  $\Gamma$  is the vertex function and the kernel  $\mathcal{K}$  is an irreducible block. The latter is calculated directly by the graph technique once the underlying dynamics is known. We should then express the vertex  $\Gamma$  through the two-body wave function. This can be done by comparing, for example, two ways of calculating the amplitude for the breakup of a bound state by some perturbation: 1) by means of the graph technique (the result contains  $\Gamma$ ); 2) by calculating the matrix element of the perturbation operator between the bound state and the free states of  $n$  particles (the result contains  $\psi$ ). We thus get:

$$\psi(k_1, k_2, p, \omega\tau) = \frac{\Gamma(k_1, k_2, p, \omega\tau)}{s - M^2}, \quad (41)$$

where  $s = (k_1 + k_2)^2 = (p + \omega\tau)^2$ . The corresponding relation for the  $n$ -body case has the same form. In any practical calculation of the amplitude, we associate  $\Gamma$  with the vertex shown in fig. 4 and then express  $\Gamma$  in terms of  $\psi$  by eq.(41).

In the simple case of a scalar particle, the equation for the wave-function in terms of the variables  $\vec{k}, \vec{n}$  has the following form:

$$(4(\vec{k}^2 + m^2) - M^2) \psi(\vec{k}, \vec{n}) = -\frac{m^2}{2\pi^3} \int \psi(\vec{k}', \vec{n}) V(\vec{k}', \vec{k}, \vec{n}, M^2) \frac{d^3 k'}{\epsilon_{k'}}. \quad (42)$$

An equation of such a type was also considered in refs. [32, 33, 34, 35, 36, 37].

In the non-relativistic limit, equation (42) turns into the Schrödinger equation in momentum space, the kernel  $V$  being the non-relativistic potential in momentum space, and the wave function no longer depends on  $\vec{n}$ .

We emphasize that the wave function, which is an equal-time wave function on the light front, turns into the ordinary wave function in the non-relativistic limit where  $c \rightarrow \infty$ . This reflects the fact that in the non-relativistic limit two simultaneous events in one frame are simultaneous in all other frames.

In the variables  $\vec{R}_\perp$  and  $x$ , eq.(42) can be rewritten in the form:

$$\left( \frac{\vec{R}_\perp^2 + m^2}{x(1-x)} - M^2 \right) \psi(\vec{R}_\perp, x) = -\frac{m^2}{2\pi^3} \int \psi(\vec{R}'_\perp, x') V(\vec{R}'_\perp, x'; \vec{R}_\perp, x, M^2) \frac{d^3 R'_\perp dx'}{2x'(1-x')} . \quad (43)$$

In this form, this equation is nothing else than the Weinberg equation [26].

The advantages of the equation for the wave function in the form (42) compared with (43) are its similarity to the non-relativistic Schrödinger equation in momentum space, and its simplicity in the case of particles with spin. These properties make eq.(42) very convenient for practical calculations.

The kernel of eq.(42) depends on the vector variable  $\vec{n}$ . We shall see that this dependence, especially the part which depends on  $M^2$ , is associated with the retardation of the interaction. From this point of view, the dependence of the wave function  $\psi(\vec{k}, \vec{n})$  on  $\vec{n}$  is a consequence of retardation.

#### 4.4 The Wick-Cutkosky model

As a simple example, we shall derive in this section the light-front wave function of a system consisting of two scalar particles with mass  $m$  interacting through the exchange of a massless scalar particle. The kernel is calculated in the ladder approximation. This is the so-called Wick-Cutkosky model. The diagrams that determine the kernel are shown in fig. 2. The kernel  $\mathcal{K}$  is given by eq.(22) with  $\mu = 0$ . Going over from the kernel  $\mathcal{K}$  to  $V = -\mathcal{K}/(4m^2)$ , introducing the constant  $\alpha = g^2/(16\pi m^2)$ , and expressing (22) by means of the initial and final relative momenta  $\vec{k}, \vec{k}'$ , we obtain [38]:

$$V = -4\pi\alpha/\vec{K}^2, \quad (44)$$

where

$$\vec{K}^2 = (\vec{k}' - \vec{k})^2 - (\vec{n} \cdot \vec{k}')(\vec{n} \cdot \vec{k}) \frac{(\varepsilon_{k'} - \varepsilon_k)^2}{\varepsilon_{k'} \varepsilon_k} + (\varepsilon_{k'}^2 + \varepsilon_k^2 - \frac{1}{2}M^2) \left| \frac{\vec{n} \cdot \vec{k}'}{\varepsilon_{k'}} - \frac{\vec{n} \cdot \vec{k}}{\varepsilon_k} \right|. \quad (45)$$

For  $k, k' \ll m$ , eq.(44) turns into the Coulomb potential in momentum space

$$V(\vec{k}', \vec{k}) \simeq -\frac{4\pi\alpha}{(\vec{k}' - \vec{k})^2}. \quad (46)$$

For  $\alpha \ll 1$ ,  $|\varepsilon_b| = |M - 2m| = m\alpha^2/4 \ll m$ , the wave function is concentrated in the non-relativistic region of momenta. The non-relativistic wave function of the ground state in the Coulomb potential has the form:

$$\psi(\vec{k}) = \frac{8\sqrt{\pi m \kappa}^{5/2}}{(\vec{k}^2 + \kappa^2)^2}, \quad (47)$$

where  $\kappa = \sqrt{m|\varepsilon_b|} = m\alpha/2$ . It is normalized, however, according to (38) with  $\varepsilon_k \approx m$  and  $N_2 = 1$ . The integral over  $d^3 k'$  in (42) is concentrated in the region  $k' \approx \kappa$ . Therefore, at  $k \gg \kappa$  the momentum  $\vec{k}'$  in  $V(\vec{k}', \vec{k}, \vec{n}, M^2)$  can be ignored, and from (42) we find:

$$\psi(\vec{k}, \vec{n}) = -\frac{mV(0, \vec{k}, \vec{n}, M^2)}{(2\pi)^3(\vec{k}^2 + \kappa^2)} \int \psi(\vec{k}') d^3 k'. \quad (48)$$

Substituting in the r.h.s. of eq.(48) the expressions (44,45) for  $V$  and (47) for  $\psi$ , we obtain

$$\psi(\vec{k}, \vec{n}) = \frac{8\sqrt{\pi m \kappa}^{5/2}}{(\vec{k}^2 + \kappa^2)^2 \left( 1 + \frac{|\vec{n} \cdot \vec{k}|}{\varepsilon_k} \right)}. \quad (49)$$

This relativistic wave function of the ground state with zero total angular momentum is a good approximation of a more exact one in the range  $k > \kappa$ . Corrections of order  $\alpha \log(\alpha)$  should be

considered in the range  $k < \kappa$  (see [39]). Though the kernel (44), (45) contains the modulus  $|\vec{n} \cdot \vec{k}'/\varepsilon_{k'} - \vec{n} \cdot \vec{k}/\varepsilon_k|$ , one can show that the exact solution of (42) has no “cusp” at  $\vec{n} \cdot \vec{k} = 0$ . This cusp in (49) appears due to our approximations.

One can check in this simple example that it is the retardation of the interaction that is the dynamical reason for the dependence of the wave function on the variable  $\vec{n}$ . The non-relativistic Coulomb expression for the kernel (46) does not contain retardation and does not depend on  $\vec{n}$  while the relativistic kernel (44) contains retardation and depends on  $\vec{n}$ . This leads to the dependence of the wave function on the argument  $\vec{n}$ .

The retardation leads to both the  $\vec{n}$ -dependence and the presence of the carriers of the interaction in the intermediate state, which contribute to the many body sectors. However, these two effects, being important in full measure in a truly relativistic system, can manifest themselves in a different way in weakly bound systems. Neglecting the many-body sectors does not necessarily entails to neglect the  $\vec{n}$ -dependence of the wave function at  $k \approx m$ . It is necessary to take into account the  $\vec{n}$ -dependence of the wave function even when one restricts to the two-nucleon sector.

We emphasize that the dependence of the wave function (49) on  $\vec{n}$  does not mean any violation of the rotational invariance. As explained above, it reflects the dependence (unavoidable one, in the field-theoretical framework) of any off-energy shell amplitude on the orientation of the light-front plane. At the same time, the on-shell amplitude expressed through the wave function should not depend on  $\vec{n}$ . For the case of electromagnetic form factor this property is discussed below in sect. 5.

The wave function of the 2p state can be found analogously. In the system where  $\vec{k}_1 + \vec{k}_2 = 0$  it has the form [38]:

$$\psi^\lambda(\vec{k}, \vec{n}) = \frac{8\pi\kappa^{7/2}m^{1/2}}{\sqrt{6}} \frac{1}{\left(\vec{k}^2 + \frac{1}{4}\kappa^2\right)^3 \left(1 + \frac{|\vec{n} \cdot \vec{k}|}{\varepsilon_k}\right)^2} \times \left\{ kY_{1\lambda}(\vec{k}/k) + Y_{1\lambda}(\vec{n}) \left[ \frac{(2\varepsilon_k - M)^2}{4\varepsilon_k M} (\vec{n} \cdot \vec{k}) - \frac{(\vec{k}^2 + \frac{1}{4}\kappa^2)}{2m} \left( \theta(-\vec{n} \cdot \vec{k}) - \theta(\vec{n} \cdot \vec{k}) \right) \right] \right\}. \quad (50)$$

The wave function corresponding to the angular momentum  $l = 1$  contains the spherical function  $Y_{1\lambda}(n)$ . This is an illustration of the fact that the vector  $\vec{n}$  participates in the construction of the total angular momentum on the same ground as the relative momentum  $\vec{k}$ . The dynamical difference between the solution with  $\vec{k} \parallel \vec{n}$  and  $\vec{k} \perp \vec{n}$  is obviously related to the property that some of the components of the angular momentum  $\vec{J}$ , before using the angular condition, depend on the interaction.

#### 4.5 Relation with the Bethe-Salpeter function

It is instructive to compare the solution (49) with one found using the Bethe-Salpeter function.

First, we find the relation between the light-front wave function and the Bethe-Salpeter function. We should start from the integral that restricts the variation of the arguments of the Bethe-Salpeter function to the light-front plane:

$$I = \int d^4x_1 d^4x_2 \delta(\omega \cdot x_1) \delta(\omega \cdot x_2) \Phi(x_1, x_2, p) \exp(ik_1 \cdot x_1 + ik_2 \cdot x_2), \quad (51)$$

where  $k_1, k_2$  are the on-shell momenta:  $k_1^2 = k_2^2 = m^2$ , and  $\Phi(x_1, x_2, p)$  is the Bethe-Salpeter function [27], eq.(21). We represent the  $\delta$ -functions in (51) by the integral form

$$\delta(\omega \cdot x) = \frac{1}{2\pi} \int \exp(-i\omega \cdot x \alpha) d\alpha,$$

introduce the Fourier transform of the Bethe-Salpeter function  $\Phi(k, p)$ ,

$$\Phi(x_1, x_2, p) = (2\pi)^{-3/2} \exp[-ip \cdot (x_1 + x_2)/2] \tilde{\Phi}(x, p), \quad x = x_1 - x_2,$$

$$\Phi(l, p) = \int \tilde{\Phi}(x, p) \exp(i l \cdot x) d^4 x ,$$

where  $l = (l_1 - l_2)/2$ ,  $p = l_1 + l_2$ ,  $l_1$  and  $l_2$  are off-mass shell four-vectors, and make the change of variables  $\alpha_1 + \alpha_2 = \tau$ ,  $(\alpha_2 - \alpha_1)/2 = \beta$ .

On the other hand, the integral (51) can be expressed in terms of the two-body light-front wave function. We assume that the light-front plane is the limit of a space-like plane, therefore the operators  $\varphi(x_1)$  and  $\varphi(x_2)$  commute, and, hence, the symbol of the  $T$  product in (21) can be omitted. In the considered representation, the Heisenberg operators  $\varphi(x)$  in (21) are identical on the light front  $\omega \cdot x = 0$  to the Schrödinger operators (just as in the ordinary formulation of field theory the Heisenberg and Schrödinger operators are identical for  $t = 0$ ). The Schrödinger operator  $\varphi(x)$  (for the spinless case for simplicity), which for  $\omega \cdot x = 0$  is the free field operator, is given by (8). We represent the state vector  $|p\rangle \equiv \phi(p)$  in (21) in the form of the expansion (26). Since the vacuum state on the light front is always “bare”, the creation operator, applied to the vacuum state  $\langle 0|$  gives zero, and in the operators  $\varphi(x)$  the part containing the annihilation operators only survives. This cuts out the two-body Fock component in the state vector. We thus obtain:

$$I = \frac{(2\pi)^{3/2}(\omega \cdot p)}{2(\omega \cdot k_1)(\omega \cdot k_2)} \int_{-\infty}^{+\infty} \psi(k_1, k_2, p, \omega\tau) \delta^{(4)}(k_1 + k_2 - p - \omega\tau) d\tau . \quad (52)$$

Comparing (51) and (52), we find:

$$\psi(k_1, k_2, p, \omega\tau) = \frac{(\omega \cdot k_1)(\omega \cdot k_2)}{\pi(\omega \cdot p)} \int_{-\infty}^{+\infty} \Phi(l_1 = k_1 - \omega\tau/2 + \omega\beta, l_2 = k_2 - \omega\tau/2 - \omega\beta, p) d\beta \quad (53)$$

where  $\Phi(l_1, l_2)$  is the Bethe-Salpeter function parametrized in terms of the off-mass shell momenta  $l_1, l_2$ . The argument  $p$  in (53) is related to the on-shell momenta  $k_1, k_2$  as  $p = k_1 + k_2 - \omega\tau$ , in contrast to off-mass shell relation  $p = l_1 + l_2$ .

In ordinary LFD, eq.(53) corresponds to the integration over  $dk_-$ . This equation makes the link between the Bethe-Salpeter function  $\Phi$  and the wave function  $\psi$  defined on the light front specified by  $\omega$ . It should be noticed however that eq.(53) is not necessarily an exact solution of eq.(42), since, as a rule, different approximations are made for the Bethe-Salpeter kernel and for the light-front one. In the ladder approximation, for example, the Bethe-Salpeter amplitude contains the box diagram, including the time-ordered diagram with two exchanged particles in the intermediate state, as indicated in graphically in eq. (76) in sect 6. This contribution is absent in the light-front ladder kernel.

Note also the interesting paper [40], (for earlier studies see [41]), where the Markov-Yukawa transversality principle for the two-body Bethe-Salpeter kernel was formulated on the covariant light-front plane. It allows not only to obtain an exact three-dimensional reduction of the Bethe-Salpeter equation, but also to make the exact reconstruction of the four-dimensional Bethe-Salpeter equation from the three-dimensional form. The three-dimensional form is convenient for spectroscopical calculations, the four-dimensional form facilitates the evaluation of the loop integrals for the form factors. In particular cases the method gives the same results as obtained earlier by other description [42, 43]. A three-quark generalization is given in [44].

The quasipotential type equations for the light-front wave function derived by restricting arguments of the Bethe-Salpeter amplitude to the light-front plane  $z + t = 0$  and corresponding electromagnetic form factors were studied in refs. [45, 46].

## 4.6 Solution in the Bethe-Salpeter approach

The exact expression for the Bethe-Salpeter function in the Wick-Cutkosky model is found in the form of the integral representation [47, 28] and, for zero angular momentum, reads:

$$\Phi(l, p) = -\frac{i}{\sqrt{4\pi}} \int_{-1}^{+1} \frac{g(z, M) dz}{(m^2 - M^2/4 - l^2 - zp \cdot l - i\epsilon)^3} . \quad (54)$$



The spectral function  $g(z, M)$  is determined by a differential equation [47, 28] and has no singularity at  $z = 0$ . The approximate explicit solution found in [47] for  $g(x, M)$  has the form:

$$g(z, M) = 2^6 \pi \sqrt{m} \kappa^{5/2} (1 - |z|) . \quad (55)$$

The discontinuity of the spectral function  $g(z, M)$  at  $z = 0$  is a result of approximation, since the solution (55) corresponds to an asymptotically small binding energy. Inserting (55) in (54) and integrating over  $z$ , one can recover the solution of the Bethe-Salpeter equation:

$$\Phi(k, p) = -ic \left[ \left( m^2 - \frac{1}{2} M^2 - k^2 \right) \left( m^2 - \left( \frac{1}{2} p + k \right)^2 - i0 \right) \left( m^2 - \left( \frac{1}{2} p - k \right)^2 - i0 \right) \right]^{-1}, \quad (56)$$

where  $c = 2^5 \sqrt{\pi m \kappa^5}$  with  $\kappa = \sqrt{m|\epsilon_b|} = m\alpha/2$ .

To find the light-front wave function, one can substitute in eq.(53) the Bethe-Salpeter function either in the form (54) or in (56). From (54) we find [48]:

$$\psi = \frac{g(1 - 2x, M)}{2^5 \sqrt{\pi} x(1 - x)(\vec{k}^2 + \kappa^2)^2} . \quad (57)$$

Substituting (55) in (57), we reproduce the expression (49) for the light-front wave function.

#### 4.7 Including spin

As explained in sect. 2.3, in the standard version of LFD the generators of the Poincaré group corresponding to the Lorentz boosts changing the orientation of the plane  $t + z = 0$ , are the dynamical ones and contain the interaction. In the explicitly covariant version of LFD the dependence of the wave function on the light-front orientation is taken into account by means of the variable  $\omega$ . Now, using kinematics (i.e., the transformation properties) we have to ensure that this wave function corresponds to a definite total angular momentum. In the case of the zero angular momentum the four-vector  $\omega$  enters always in the scalar product with the particle four-momenta. For the non-zero spins  $\omega$  appears in the spin structures.

We illustrate the construction of the states with spins by two examples.

Consider a system consisting of quark and antiquark in the  $J^\pi = 0^-$  state ("pion"). The light-front wave function has the form:

$$\psi = \bar{u}(k_2) \left[ A_1 \frac{1}{m} + A_2 \frac{\hat{\omega}}{\omega \cdot p} \right] \gamma_5 v(k_1), \quad (58)$$

where  $\bar{u}$  and  $v$  are the spinors,  $\hat{\omega} = \omega_\mu \gamma^\mu$ ,  $A_{1,2}$  are the scalar functions,  $m$  is the quark mass. In the system of reference where  $\vec{k}_1 + \vec{k}_2 = 0$  this wave function obtains the form:

$$\psi = w_2^t \left( g_1 + \frac{i \vec{\sigma} \cdot [\vec{n} \times \vec{k}]}{k} g_2 \right) w_1, \quad (59)$$

with the following relations between the invariant functions:

$$A_1 = -\frac{m}{2\varepsilon_k} (g_1 + \frac{m}{k} g_2), \quad A_2 = \frac{\varepsilon_k}{k} g_2 .$$

Note that there exists a special representation (see [11]) in which the wave function has the form (59) in arbitrary system of reference.

From eqs.(58,59) one can see that the spin structure of the wave function indeed contains the four-vector  $\omega$  determining the light-front orientation. Due to that it is determined by two invariant functions. Only one of them ( $g_1$ ) survives in the nonrelativistic limit.

Another example is the light-front wave function of a system consisting of two fermions in the state with total angular momentum equal to 1. This can be two nucleons in the state  $J^\pi = 1^+$

(the deuteron) or the quark-antiquark pair in the state  $J^\pi = 1^-$  ( $\rho$ -meson). This wave function has the form:

$$\Phi_{\sigma_2\sigma_1}^\lambda(k_1, k_2, p, \omega\tau) = \sqrt{m}e_\mu^\lambda(p)\bar{u}^{\sigma_2}(k_2)\phi_\mu U_c \bar{u}^{\sigma_1}(k_1), \quad (60)$$

with

$$\begin{aligned} \phi_\mu = & \varphi_1 \frac{(k_1 - k_2)_\mu}{2m^2} + \varphi_2 \frac{1}{m} \gamma_\mu + \varphi_3 \frac{\omega_\mu}{\omega \cdot p} + \varphi_4 \frac{(k_1 - k_2)_\mu \hat{\omega}}{2m\omega \cdot p} \\ & - \varphi_5 \frac{i}{m^2 \omega \cdot p} \gamma_5 \epsilon_{\mu\nu\rho\gamma} k_{1\nu} k_{2\rho} \omega_\gamma + \varphi_6 \frac{m\omega_\mu \hat{\omega}}{(\omega \cdot p)^2}. \end{aligned} \quad (61)$$

It is determined by six invariant functions  $\varphi_{1-6}$ , depending on two scalar variables. This number is the dimension of the matrix depending on the spin projections of the deuteron and two nucleons, divided by the factor 2 due to the parity conservation:  $N = 3 \times 2 \times 2/2 = 6$ .

In the system of reference where  $\vec{k}_1 + \vec{k}_2 = 0$  (or in arbitrary system, but in the representation described in [11]) this wave function obtains the form:

$$\Psi_{\sigma_2\sigma_1}^\lambda(\vec{k}, \vec{n}) = \sqrt{m}w_{\sigma_2}^\dagger \psi^\lambda(\vec{k}, \vec{n}) \sigma_y w_{\sigma_1}^\dagger, \quad (62)$$

with

$$\begin{aligned} \vec{\psi}(\vec{k}, \vec{n}) = & f_1 \frac{1}{\sqrt{2}} \vec{\sigma} + f_2 \frac{1}{2} \left( \frac{3\vec{k}(\vec{k} \cdot \vec{\sigma})}{\vec{k}^2} - \vec{\sigma} \right) + f_3 \frac{1}{2} (3\vec{n}(\vec{n} \cdot \vec{\sigma}) - \vec{\sigma}) \\ & + f_4 \frac{1}{2k} (3\vec{k}(\vec{n} \cdot \vec{\sigma}) + 3\vec{n}(\vec{k} \cdot \vec{\sigma}) - 2(\vec{k} \cdot \vec{n})\vec{\sigma}) \\ & + f_5 \sqrt{\frac{3}{2}} \frac{i}{k} [\vec{k} \times \vec{n}] + f_6 \frac{\sqrt{3}}{2k} [[\vec{k} \times \vec{n}] \times \vec{\sigma}], \end{aligned} \quad (63)$$

where  $w$  is the two-component nucleon spinor normalized to  $w^\dagger w = 1$ . The relations between  $\varphi$  and  $f$  can be found in [11]. In the relativistic one boson exchange model this wave function was calculated in [49]. It was found that the function  $f_5$ , of relativistic origin, is very important: it dominates at  $k > 500$  MeV/c. In nonrelativistic the functions  $f_{3-6}$  become negligible, and only two first structures survive, corresponding to usual S- and D-waves.

This wave function was used in the paper [50] to calculate the deuteron electromagnetic form factors. No any parameters were fitted. It turned out that the calculated structure function  $A(Q^2)$  and the polarization observable  $t_{20}$  coincide with rather precise experimental data obtained recently at CEBAF/TJNAF.

#### 4.8 The nucleon wave function

Many calculations of the nucleon properties (magnetic moments, form factors, etc.) are carried out in the framework of LFD with the nucleon wave function in the  $3q$  model containing one or a few spin components. The total number of the spin components in the nucleon wave function is sixteen [51]. This is related to the fact known long ago [52] that in a many-body system the parity conservation does not reduce the number of the spin components. This is so for a relativistic three-body system and for any  $n$ -body system for  $n \geq 4$  (both relativistic and nonrelativistic one). Hence, for the relativistic nucleon we get

$$N = (2S_1 + 1)(2S_2 + 1)(2S_3 + 1)(2S_N + 1) = 2 \times 2 \times 2 \times 2 = 16.$$

These 16 components are forming the full basis for the nucleon wave function.

In nonrelativistic limit the parity conservation reduces this number down to 8. Their relativistic counterparts were found in [53]. Note, however, that one can construct also another 8 components with the opposite parity.

The difference between relativistic and nonrelativistic cases is related to the fact that in relativistic case one can construct the pseudoscalar:

$$C_{ps} = \epsilon^{\mu\nu\rho\gamma} k_{1\mu} k_{2\nu} k_{3\rho} p_\gamma \quad (64)$$

It is not zero, since the bound quarks are off-energy-shell:  $k_1 + k_2 + k_3 = p + \omega\tau \neq p$ . In ordinary light-front approach this corresponds to the well known conservation law:

$$\vec{k}_{1\perp} + \vec{k}_{2\perp} + \vec{k}_{3\perp} = \vec{p}_\perp, \quad k_{1+} + k_{2+} + k_{3+} = p_+,$$

but  $k_{1-} + k_{2-} + k_{3-} \neq p_-$ . Therefore, we can take 8 componets with opposite parity, multiply them by  $C_{ps}$  and get another 8 componets with the nucleon parity. By this way, we get 16 components of the nucleon wave function. They are given in [51]. Due to the momentum conservation, the pseudoscalar (64) can be rewritten as:

$$C_{ps} = -\tau\epsilon^{\mu\nu\rho\gamma} k_{1\mu} k_{2\nu} k_{3\rho} \omega_\gamma.$$

It is proportional to  $\omega$ . So, namely the dependence of the relativistic nucleon wave function on the light-front orientation  $\omega$  is the reason of appearance of 8 extra componets. In nonrelativistic case this dependence disappears, and we remain with 8 components. Formally, this is due to the fact, that  $\omega$  enters in the momentum conservation law in the combination  $\omega\tau$ , where  $\tau = (s - M^2)/(2\omega \cdot p)$ . This term contains extra factor  $k/m$  and disappears at  $k \ll m$ . We get the nonrelativistic conservation law:  $\vec{k}_1 + \vec{k}_2 + \vec{k}_3 = \vec{p}$  and loose opportunity to construct any pseudoscalar and the extra components.

As mentioned above, an advantage of the explicitly covariant LFD is simplification of the transformation properties of the wave functions with a given spin. In the standard LFD approach the wave function is transformed in every spin index by a special Melosh rotation matrices [54]. In the covariant version the transformation properties are automatically taken into account and does not require any Melosh matrices.

Consider, for example, the nucleon wave function in c.m.-system with fully symmetrical S-wave spin-isospin structure:

$$\Psi_S = \frac{\psi_S}{\sqrt{72}} [3 + (\vec{\sigma}_{12} \cdot \vec{\sigma}_{3N})(\vec{\tau}_{12} \cdot \vec{\tau}_{3N})] \quad (65)$$

where  $\vec{\sigma}_{12} = (w_1^\dagger \vec{\sigma} w_2)$ ,  $\vec{\sigma}_{3N} = (w_3^\dagger \vec{\sigma} w_N)$  and similarly for the isospin matrices  $\vec{\tau}_{12}$ ,  $\vec{\tau}_{3N}$ . Using the Fierz identities, one can check that the wave function (65) is indeed symmetric relative to permutations (provided  $\psi_S$  is symmetric). In arbitrary system it is multiplied by the Melosh rotation matrices. For  $\psi_S$  one can take, for example, the harmonic oscillator model:

$$\psi_S = \frac{2^4 \pi^{3/2} 3^{1/4} N}{\alpha^3} \exp \left( -\frac{\vec{k}_1^2 + \vec{k}_2^2 + \vec{k}_3^2}{2\alpha^2} \right),$$

$\vec{k}_i$  are the quark relative momenta,  $N$  is a normalisation factor equal to 1 in the nonrelativistic limit.

In the explicitly covariant LFD it is represented in covariant, four-dimensional form, in terms of the usual Dirac spinors, avoiding any Melosh matrices. For this aim we introduce the projection operators:

$$\Pi_+ = \frac{\mathcal{M} + \hat{\mathcal{P}}}{2\mathcal{M}}, \quad \Pi_- = \frac{\mathcal{M} - \hat{\mathcal{P}}}{2\mathcal{M}},$$

where  $\mathcal{P} = k_1 + k_2 + k_3 = p + \omega\tau$ ,  $\hat{\mathcal{P}} = \gamma^\mu \mathcal{P}_\mu$ ,  $\mathcal{M}^2 = \mathcal{P}^2$ .  $\mathcal{M}$  here is the effective mass of the free quarks (not the nucleon mass). Then the wave function (65) is covariantly represented as [51]:

$$\begin{aligned} \Psi_S = & \frac{\psi_S}{\sqrt{72}} c_1 c_2 c_3 c_N \{ 3 [\bar{u}(k_1) \Pi_+ \gamma_5 U_c \bar{u}(k_2)] [\bar{u}(k_3) \Pi_+ u(p)] \\ & - [\bar{u}(k_1) \Pi_+ \gamma^\mu \Pi_- U_c \bar{u}(k_2)] [\bar{u}(k_3) \Pi_+ \gamma_\mu \gamma_5 \Pi_+ u(p)] (\vec{\tau}_{12} \cdot \vec{\tau}_{3N}) \}, \end{aligned} \quad (66)$$

where  $c_{1,2,3} = 1/\sqrt{\epsilon_{1,2,3} + m}$ ,  $c_N = 1/\sqrt{\epsilon_N + M}$  and, e.g.,  $\epsilon_1 = \sqrt{\vec{k}_1^2 + m^2}$  is the energy of the quark 1. In the system where  $\vec{k}_1 + \vec{k}_2 + \vec{k}_3 = 0$  this wave function *exactly* coincides with (65). The wave function (66) can be decomposed in terms of the 16 structures discussed above. Other

states are represented similarly. The calculation of the nucleon properties (magnetic moments, electromagnetic form factors, etc.) is now a standard routine using the trace techniques of the Dirac matrices. In comparison to the standard light-front approach, for the identical nucleon wave functions, the results in both approaches coincide with each other, but in the explicitly covariant approach they are obtained much more simpler.

## 5 Electromagnetic form factors

The general physical electromagnetic amplitude of a spinless system is given by:

$$J_\rho \equiv \langle p' | J_\rho(0) | p \rangle = (p + p')_\rho F(Q^2) . \quad (67)$$

where  $F(Q^2)$  is the electromagnetic form factor. In LFD it is obtained by calculating the amplitude corresponding to fig. 6:

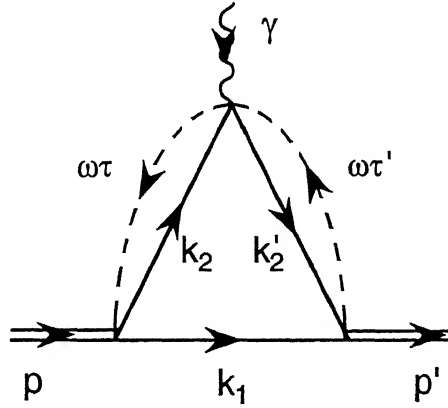


Fig.6. Electromagnetic vertex of a bound system.

$$J_\rho = \frac{1}{(2\pi)^3} \int \frac{(p + p' + \omega\tau + \omega\tau' - 2k_1)_\rho}{(1 - \omega \cdot k_1 / \omega \cdot p)^2} \psi' \psi \theta(\omega \cdot (p - k_1)) \frac{d^3 k_1}{2\varepsilon_{k_1}} . \quad (68)$$

Exact light-front amplitude on the energy shell has to coincide with the Feynman one and should not depend on the orientation of the light front plane. It should reproduce the form (67). However, the diagram 6 corresponds to impulse approximation, when the electromagnetic current does not contain any interaction. Therefore the dependence of amplitude (68) on the light-front orientation survives.  $J_\rho$  depends on  $\omega$ . It also can be represented in the general form:

$$J_\rho = (p + p')_\rho F(Q^2) + \frac{\omega_\rho}{\omega \cdot p} B_1(Q^2) . \quad (69)$$

The factor  $1/\omega \cdot p$  is separated for convenience. The invariant functions  $F$  and  $B_1$  depend on  $Q^2 = -q^2 \equiv -(p' - p)^2$ . They could depend in principle on  $\omega \cdot p$  and  $\omega \cdot p'$ . However, the four-vector  $\omega$  is defined up to an arbitrary number, and, hence, the theory is invariant relatively to the replacement  $\omega \rightarrow \alpha\omega$ , where  $\alpha$  is a number. The form factors  $F$  and  $B_1$  can therefore depend only on the ratio  $\omega \cdot p' / \omega \cdot p$ .

Now we take into account that  $\omega$  is restricted by the condition  $\omega \cdot q = 0$ , implying the transversality of  $q$ . In this case we have  $\omega \cdot p' / \omega \cdot p = 1$ , and the functions  $F$  and  $B_1$  depend on  $Q^2$  only.

The main difference of the amplitude (69) with respect to (67) is the presence of an additional contribution, proportional to  $\omega_\rho$ . To avoid any misunderstanding, we emphasize that even *in the case where the wave function  $\psi$  does not depend on  $\vec{n}$* , the term proportional to  $\omega_\rho$  still survives in the electromagnetic vertex.

In the spinless case, the physical form factor,  $F(Q^2)$  can be obtained immediately by multiplying both sides of eq.(69) by  $\omega_\rho$ . We thus get:

$$F(Q^2) = \frac{J \cdot \omega}{2\omega \cdot p}. \quad (70)$$

With (68), (70) we obtain:

$$F(Q^2) = \frac{1}{(2\pi)^3} \int \psi(\vec{R}_\perp^2, x) \psi((\vec{R}_\perp - x\vec{\Delta})^2, x) \frac{d^2 R_\perp dx}{2x(1-x)}. \quad (71)$$

We have represented here, and in the following, the four-momentum transfer  $q$  by  $q = (q_0, \vec{\Delta}, \vec{q}_\parallel)$  with  $\vec{\Delta} \cdot \vec{\omega} = 0$  and  $\vec{q}_\parallel$  is parallel to  $\vec{\omega}$ . Since  $\omega \cdot q = 0$ , we have  $Q^2 = -q^2 = \vec{\Delta}^2$ .

The form factor in the Bethe-Salpeter approach is found from the formula:

$$(p + p')_\rho F(t) = i \int (p + p' - 2k)_\rho \Phi(\frac{1}{2}p - k, p) \Phi(\frac{1}{2}p' - k, p') (m^2 - k^2) \frac{d^4 k}{(2\pi)^4}. \quad (72)$$

The Bethe-Salpeter function  $\Phi(k, p)$  is given by eq. (56). *The form factors calculated by means of both approach coincide with each other with high accuracy.* Both approaches give the same asymptotical behavior of the form factors at  $|t| \gg m^2$ :

$$F(t) \approx \frac{16\alpha^4 m^4}{t^2} \left[ 1 + \frac{\alpha}{2\pi} \log \left( \frac{|t|}{m^2} \right) \right],$$

where  $\alpha = g^2/(16\pi m^2)$ ,  $g$  is the coupling constant in the Wick-Cutkosky model.

In the usual light-front formulation, with  $\omega = (1, 0, 0, -1)$ , eq.(70) corresponds to expressing the form factor through the  $J_+$  component. This is well known, and eq.(71) has been found in ref. [55]. However, this procedure cannot be extended to the calculation of physical form factors of systems with total spin 1/2 and 1. Their electromagnetic vertices also depend on the four-vector  $\omega$ .

For for a spin-1 particle this vertex has the form:

$$\langle \lambda' | J_\rho | \lambda \rangle = \frac{1}{2\omega \cdot p} e_\mu^{*\lambda'}(p') J_\rho^{\mu\nu} e_\nu^\lambda(p), \quad \text{where} \quad J_\rho^{\mu\nu} = T_\rho^{\mu\nu} + B_\rho^{\mu\nu}(\omega). \quad (73)$$

Here  $T_\rho^{\mu\nu}$  is determined by the physical form factors and has the usual structure [56]:

$$\begin{aligned} \langle \lambda' | J_\rho | \lambda \rangle &= e_\mu^{*\lambda'}(p') \left\{ P_\rho \left[ \mathcal{F}_1(q^2) g^{\mu\nu} + \mathcal{F}_2(q^2) \frac{q^\mu q^\nu}{2M^2} \right] + \mathcal{G}_1(q^2) (g_\rho^\mu q^\nu - g_\rho^\nu q^\mu) \right\} e_\nu^\lambda(p) \\ &\equiv e_\mu^{*\lambda'}(p') T_\rho^{\mu\nu} e_\nu^\lambda(p), \end{aligned} \quad (74)$$

$e_\mu^\lambda(p)$  is the spin-1 polarization vector,  $p$  and  $p'$  are the initial and final momenta,  $\lambda$  and  $\lambda'$  are the corresponding helicities,  $P = p + p'$  and  $q = p' - p$ . The tensor  $B_\rho^{\mu\nu}$  contains the  $\omega$  dependent terms:

$$\begin{aligned} B_\rho^{\mu\nu} &= \frac{M^2}{2(\omega \cdot p)} \omega_\rho \left[ B_1 g^{\mu\nu} + B_2 \frac{q^\mu q^\nu}{M^2} + B_3 M^2 \frac{\omega^\mu \omega^\nu}{(\omega \cdot p)^2} + B_4 \frac{q^\mu \omega^\nu - q^\nu \omega^\mu}{2\omega \cdot p} \right] \\ &+ B_5 P_\rho M^2 \frac{\omega^\mu \omega^\nu}{(\omega \cdot p)^2} + B_6 P_\rho \frac{q^\mu \omega^\nu - q^\nu \omega^\mu}{2\omega \cdot p} + B_7 M^2 \frac{g_\rho^\mu \omega^\nu + g_\rho^\nu \omega^\mu}{\omega \cdot p} \\ &+ B_8 q_\rho \frac{q^\mu \omega^\nu + q^\nu \omega^\mu}{2\omega \cdot p}, \end{aligned} \quad (75)$$

$B_1, \dots, B_8$  are invariant functions. This tensor is not eliminated by contraction with  $\omega_\rho$ . In these cases the electromagnetic form factors are given by contraction of the electromagnetic vertex with more complicated tensors found in [57, 58]. The current component  $J_+$  is still enough to find the form factors  $\mathcal{F}_1, \mathcal{F}_2$ , but it is not enough to find  $\mathcal{G}_1$ .

The formulas for the physical form factors for the case of spin-1/2 light-front electromagnetic vertex (nucleon electromagnetic form factors, for instance) are found in [59].

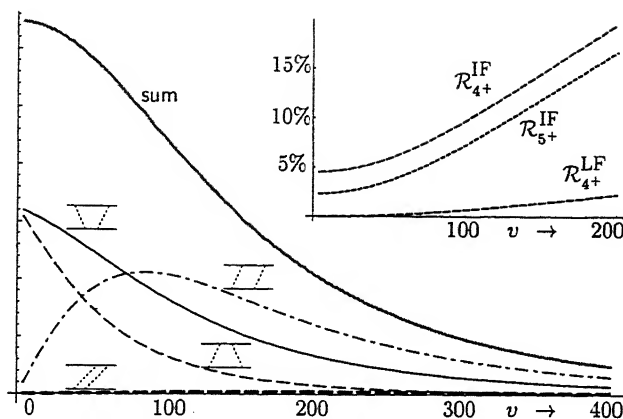


Fig.7. LF time-ordered boxes for a scattering angle of  $\pi/2$  as a function of the incoming momentum  $v$ . We also give the ratios of boxes with at least four particles ( $\mathcal{R}_{4+}^{IF}$  and  $\mathcal{R}_{4+}^{LF}$ ) or five particles ( $\mathcal{R}_{5+}^{IF}$ ,  $\mathcal{R}_{5+}^{LF} = 0$ ) in one of the intermediate states.

## 6 Suppression of the higher Fock states

The kernel corresponding to exchange by a particle in the Bethe-Salpeter approach and in LFD are not equivalent to each other. The light-front graphs are obtained from the Feynman ones by time-ordering of the vertices. For example, the Feynman graph with two exchanges corresponds to the following sum of the time ordered graphs:

$$\begin{array}{c} \text{---} \text{---} \text{---} \\ | \quad | \\ \text{---} \text{---} \end{array} = \begin{array}{c} \text{---} \text{---} \text{---} \\ \diagup \quad \diagdown \\ \text{---} \text{---} \end{array} + \begin{array}{c} \text{---} \text{---} \text{---} \\ \diagdown \quad \diagup \\ \text{---} \text{---} \end{array} + \begin{array}{c} \text{---} \text{---} \text{---} \\ \text{---} \text{---} \text{---} \\ \text{---} \text{---} \end{array} \\ + \begin{array}{c} \text{---} \text{---} \text{---} \\ \text{---} \text{---} \text{---} \\ \text{---} \text{---} \end{array} + \begin{array}{c} \text{---} \text{---} \text{---} \\ \text{---} \text{---} \text{---} \\ \text{---} \text{---} \end{array} + \begin{array}{c} \text{---} \text{---} \text{---} \\ \text{---} \text{---} \text{---} \\ \text{---} \text{---} \end{array} \quad (76)$$

The last two graphs in (76) containing two exchanged particles in the intermediate state ("the stretched box") are omitted in the second iteration of the light-front kernel. The number of graphs with increasing number of intermediate particles increases in higher orders. At small value of the coupling constant  $\alpha$  their contribution can be suppressed, but at  $\alpha \approx 1$  this reason of the suppression disappears. *However, these higher Fock state graphs are still suppressed.*

In the papers [60, 61] the binding energy was calculated in the framework of the Bethe-Salpeter equation and the light-front one. It was found that even at  $\alpha \approx 1$  the binding energies calculated in both approaches are very close to each other. This indicates that the contribution of the higher Fock states is suppressed.

This contribution has been calculated directly in the papers [62, 63]. The result is shown in fig. 7. In these figures  $v$  means the incoming momentum. One can see that the contribution of the stretched box into the sum of time ordered graphs is negligible. Its relative contribution  $\mathcal{R}_{4+}^{LF}$  is of the order a few per cent.

Another important conclusion which follows from fig. 7 is that the suppression of the higher Fock states takes place namely in LFD. In the instant form of dynamics these contributions much more larger. For four or more intermediate particles, due to the fluctuations, they are indicated in fig. 7 as  $\mathcal{R}_{4+}^{IF}$ . The corresponding graphs are shown in fig. 8. For five or more intermediate particles, due to a few vacuum vertices, they are indicated as  $\mathcal{R}_{5+}^{IF}$ .

These results show that the light-front contributions of higher Fock states are significantly smaller than in the instant form. In the limit  $v \rightarrow 0$  the ratio  $\mathcal{R}_{4+}^{LF}$  goes to zero, because the phase space becomes empty. However, in the instant form there is a finite contribution of  $\mathcal{R}_{4+}^{IF} = 4.5\%$  in this limit.

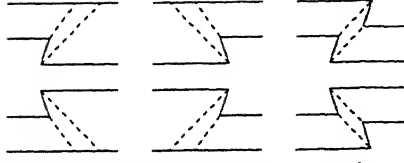


Fig.8. Time-ordered diagrams that contribute to  $\mathcal{R}_5$ . The diagrams in the first column have five particles in the first intermediate state. The diagrams in the second column have five particles in the last intermediate state, and the diagrams on the right have five-particle intermediate states for both the first and the third intermediate state.

## 7 Conclusion

We have described the general construction of LFD, its explicitly covariant formulation and some applications to the field theory and to the relativistic wave functions. These developments have been made particularly simple, and intuitive, by the three-dimensional nature of formalism, interpretation of amplitudes in terms of the space-time picture and the absence of vacuum fluctuations. We have shown also the relation to other approaches, in particular, to the Bethe-Salpeter one.

Though the light-front amplitudes can be derived from the Feynman ones, and the light-front wave function can be obtained by the projection of the Bethe-Salpeter amplitude on the light-front plane, this does not mean that LFD is only a method to calculate the Feynman amplitudes and to find an approximate eigenvalue of the Bethe-Salpeter equation.

The light-front approach has much more general and independent meaning. It is an alternative and rather powerful way to solve the field-theoretical problems.

## 8 Appendix

### 8.1 Kinematical transformations

We specify here the transformation properties of the state vector with respect to transformations of the coordinate system.

The operators associated to the four-momentum and four-dimensional angular momentum are expressed in terms of integrals of the energy-momentum  $T_{\mu\nu}$  and the angular momentum  $M_{\mu\nu}^\rho$  tensors over the light-front plane  $\omega \cdot x = \sigma$ , according to:

$$P_\mu = \int T_{\mu\nu} \omega^\nu \delta(\omega \cdot x - \sigma) d^4x = P_\mu^0 + P_\mu^{int}, \quad (77)$$

$$J_{\mu\nu} = \int M_{\mu\nu}^\rho \omega_\rho \delta(\omega \cdot x - \sigma) d^4x = J_{\mu\nu}^0 + J_{\mu\nu}^{int}, \quad (78)$$

where the 0 and *int* superscripts indicate the free and interacting parts of the operators respectively. For generality, we consider here the light-front time  $\sigma \neq 0$ .

The description of the evolution along the light-front time  $\sigma$  implies a fixed value of the length of  $\vec{\omega}$ , or, equivalently, of  $\omega_0$ . This is necessary in order to have a scale of  $\sigma$ . However, the most important properties of the physical amplitudes following from covariance do not require to fix the scale of  $\omega$  and will be invariant relative to its change. We work in the interaction representation in which the operators are expressed in terms of the free fields. Consider, for example, the scalar field  $\varphi(x)$ , eq.(8). Then the free operators  $P_\mu^0$  have the form:

$$P_\mu^0 = \int a^\dagger(\vec{k}) a(\vec{k}) k_\mu d^3k, \quad (79)$$

$$J_{\mu\nu}^0 = \int a^\dagger(\vec{k}) a(\vec{k}) i \left( k_\mu \frac{\partial}{\partial k^\nu} - k_\nu \frac{\partial}{\partial k^\mu} \right) d^3k, \quad (80)$$

The operators  $P^{int}$  and  $J^{int}$  contain the interaction Hamiltonian  $H^{int}(x)$ :

$$P_\mu^{int} = \omega_\mu \int H^{int}(x) \delta(\omega \cdot x - \sigma) d^4x, \quad (81)$$

$$J_{\mu\nu}^{int} = \int H^{int}(x) (x_\mu \omega_\nu - x_\nu \omega_\mu) \delta(\omega \cdot x - \sigma) d^4x. \quad (82)$$

The field-theoretical Hamiltonian  $H^{int}(x)$  is usually singular and requires a regularization. The regularization of amplitudes will be illustrated above in sect. 3.3 by the example of a typical self-energy contribution.

In the particular case  $\omega = (1, 0, 0, -1)$ , in the light-front coordinates, only  $\omega_-$ -component is non-zero. This just gives that in (80,82) the components  $P_-^{int}, J_{+-}^{int}$  are non-zero, i.e., corresponding generators in (77,78) contains the interaction.

Under translation  $x \rightarrow x' = x + a$  of the coordinate system  $A \rightarrow A'$ , the equation  $\omega \cdot x = \sigma$  takes the form  $\omega \cdot x' = \sigma'$ , where  $\sigma' = \sigma + \omega \cdot a$ . The state vector is transformed as:

$$\phi_\omega(\sigma) \rightarrow \phi'_\omega(\sigma') = U_{P^0}(a) \phi_\omega(\sigma), \quad (83)$$

where the operator  $U_{P^0}(a)$  contains only the operator of the four-momentum (79) of the free field:

$$U_{P^0}(a) = \exp(iP^0 \cdot a). \quad (84)$$

The “prime” at  $\phi'(\sigma)$  indicates that  $\phi'(\sigma)$  is defined in the system  $A'$  on the plane  $\omega \cdot x' = \sigma$  in contrast to  $\phi(\sigma)$  defined in the system  $A$  on the plane  $\omega \cdot x = \sigma$  (the value of  $\sigma$  being the same). The state vector  $\phi'(\sigma')$  is defined in  $A'$  on the plane  $\omega \cdot x' = \sigma'$ , which coincides with  $\omega \cdot x = \sigma$ . Therefore no dynamics enters into the transformation (83). This is rather natural, since under translation of the coordinate system the plane  $\omega \cdot x = \sigma$  occupies the same position in space while it occupies a new position with respect to the axes of the new coordinate system, as indicated in fig. 9. The formal proof of (83), (84) can be found in [64].

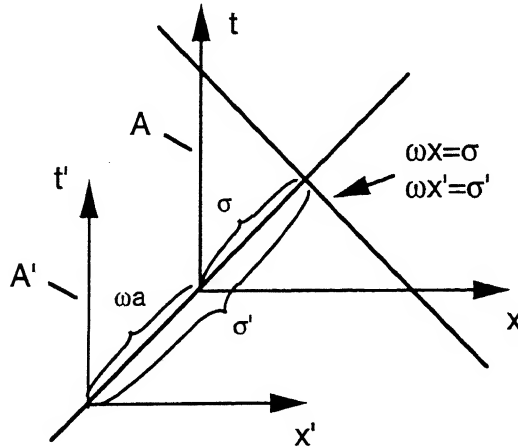


Fig.9. Translation of the reference system along the light-front time.

In the case of infinitesimal four-dimensional rotations  $x_\mu \rightarrow x'_\mu = g x_\mu = x_\mu + \epsilon_{\nu\mu} x^\nu$ , the result is similar [64]:

$$\phi_\omega(\sigma) \rightarrow \phi'_{\omega'}(\sigma) = U_{J^0}(g) \phi_\omega(\sigma), \quad (85)$$

where  $\omega'_\mu = \omega_\mu + \epsilon_{\nu\mu} \omega^\nu$  and

$$U_{J^0}(g) = 1 + \frac{1}{2} J_{\mu\nu}^0 \epsilon^{\mu\nu}. \quad (86)$$

The operator  $J_{\mu\nu}^0$  is given by (80). This shows that the transformations of the state vector with respect to the transformations of the coordinate system are indeed kinematical.



## 8.2 Dynamical transformations

The properties of the state vector under transformations of the hypersurface are determined by the dynamics and follow from the Tomonaga-Schwinger equation [65]:

$$i\delta\phi/\delta\sigma(x) = H^{int}(x) \phi . \quad (87)$$

From the definition of the variational derivative in (87) we obtain:

$$i\delta\phi = H^{int}(x) \phi \delta V(x) ,$$

where  $\delta V(x)$  is the volume between the initial surface and the surface obtained from the original one by the variation  $\delta\sigma(x)$  around the point  $x$ .

Under the translation  $\sigma \rightarrow \sigma + \delta\sigma$  of the plane, the total increment of the state vector is obtained through the increment at each point of the surface:

$$i\delta\phi = \int H^{int}(x) \delta(\omega \cdot x - \sigma) d^4x \phi \delta\sigma . \quad (88)$$

This relation gives the Schrödinger equation. In the interaction representation in the light-front time, we have:

$$i\partial\phi/\partial\sigma = H(\sigma)\phi(\sigma) , \quad (89)$$

where:

$$H(\sigma) = \int H_\omega^{int}(x) \delta(\omega \cdot x - \sigma) d^4x , \quad (90)$$

and  $H_\omega^{int}(x)$  may differ from  $H^{int}(x)$  because of singularities of the field commutators on the light cone. This point is explained below in the section 3.

Similarly, in the case of a rotation of the light-front plane,  $\omega_\mu \rightarrow \omega'_\mu = \omega_\mu + \delta\omega_\mu$ ,  $\delta\omega_\mu \approx \epsilon_{\nu\mu}\omega^\nu$ , we find:

$$\phi_\omega(\sigma) \rightarrow \phi_{\omega+\delta\omega}(\sigma) = \phi_\omega + \delta\phi_\omega , \quad \delta\phi_\omega = \frac{1}{2}\epsilon_{\mu\nu} \left( \omega^\mu \frac{\partial}{\partial\omega_\nu} - \omega^\nu \frac{\partial}{\partial\omega_\mu} \right) \phi_\omega(\sigma) . \quad (91)$$

The increment of the volume over the point  $x$  is:

$$\delta V = \epsilon_{\mu\nu} x^\mu \omega^\nu \delta(\omega \cdot x - \sigma) d^4x , \quad (92)$$

and it follows from (88) that [64]:

$$J_{\mu\nu}^{int} \phi_\omega(\sigma) = L_{\mu\nu}(\omega) \phi_\omega(\sigma) , \quad (93)$$

where:

$$L_{\mu\nu}(\omega) = i \left( \omega_\mu \frac{\partial}{\partial\omega^\nu} - \omega_\nu \frac{\partial}{\partial\omega^\mu} \right) , \quad (94)$$

and  $J_{\mu\nu}^{int}$  is given by (82).

Equation (93) is called the *angular condition*. It plays an important role in the construction of relativistic bound states.

The transformation of the coordinate system and the simultaneous transformation of the light-front plane, which is rigidly related to the coordinate axes, correspond to the successive application of the two types of transformations considered above (kinematical and dynamical). Thus, under the infinitesimal translation  $x \rightarrow x' = x + a$  of the coordinate system,  $A \rightarrow A'$ , and of the plane, we have:

$$\phi_\omega(\sigma) \rightarrow \phi'_\omega(\sigma) = (1 + iP \cdot a) \phi_\omega(\sigma) . \quad (95)$$

Note that for the state with definite total four-momentum  $p$  (i.e., for an eigenstate of the four-momentum operator), the equations (83) and (95) give:

$$\exp(iP^0 \cdot a) \phi(\sigma) = \exp(ip \cdot a) \phi(\sigma + \omega \cdot a) . \quad (96)$$

This equation determines the conservation law (27) for the four-momenta of the constituents.

## References

- [1] P.A.M. Dirac, Rev. Mod. Rhys. **21** (1949) 392.
- [2] J. Kogut and L. Suskind, Phys. Reports, **8** (1973) 75.
- [3] H. Leutwyler and J. Stern, Ann. Phys. (N.Y.) **112** (1978) 94.
- [4] F.M. Lev, Fortschr. Phys. **31** (1983) 75.
- [5] *Hadronic Physics with Multi-GeV Electrons*, eds. B. Desplanques and D. Goutte, Nova Science, Commack, New York, 1990;
- [6] B.D. Keister and W.N. Polyzou, In: *Advances in Nuclear Physics*, ed. J.W. Negele and E.W. Vogt, (Plenum Press, New York) **20** (1991) 225.
- [7] F. Coester, Prog. in Part. and Nucl. Phys., **29** (1992) 1.
- [8] V.R. Garisevanishvili and Z.R. Menteshashvili, "*Relativistic Nuclear Physics in the Light Front Formalism*", Nova Science Publishers, New York, 1993.
- [9] B.D. Keister, AIP Conf. Proc. 334, Few-Body Problems in Physics, p.164, Williamsburg, May 1994, ed. F. Gross, AIP press, New York.
- [10] "*Theory of Hadrons and LFQCD*", Fourth International Workshop on Light-Front Quantization and Non-Perturbative Dynamics, August 1994, ed. St. Glazek, World Scientific, 1995.
- [11] J. Carbonell, B. Desplanques, V.A. Karmanov and J.-F. Mathiot, Phys. Reports, **300** (1998) 215.
- [12] F.M. Lev, Nucl. Phys. **A606** (1996) 459.
- [13] *New Non-Perturbative Methods and Quantization on the Light Cone*, Les Houches School, Feb. 24 - March 7, 1997, v. 8, ed. by P. Grangé et al., Springer-EDP Sciences, 1998.
- [14] St. Glazek, Acta Phys. Polon., **B29** (1998) 3558.
- [15] M.M. Brisudova, R.J. Perry and K.G. Wilson, Phys. Rev. Lett., **78** (1997) 1227.
- [16] S.J. Brodsky, R. Roskies and R. Suaya, Phys. Rev. **D8** (1973) 4574.
- [17] G.P. Lepage and S.J. Brodsky, Phys. Rev. **D22** (1980) 2157.
- [18] V.A. Karmanov, ZhETF, **71** (1976) 399 [transl.: JETP, **44** (1976) 210].
- [19] F. Coester, W.H. Klink and W.N. Polyzou, Few-Body Syst. Suppl. **10** (1999) 115.
- [20] V.G. Kadyshevsky, ZhETF, **46** (1964) 654, 872 [JETP, **19** (1964) 443, 597].
- [21] V.G. Kadyshevsky, Nucl. Phys. **B6** (1968) 125.
- [22] V.G. Kadyshevsky, R.M. Mir-Kasimov and N.B. Skachkov, Fiz. Elem. Chastits At. Yadra, **2** (1972) 635 [Sov. J. Part. Nucl. **2** (1972) 69].
- [23] V.A. Karmanov, Fiz. Elem. Chastits At. Yadra, **19** (1988) 525 [Sov. J. Part. Nucl. **19** (1988) 228].
- [24] N.E. Ligterink and B.L.G. Bakker, Phys. Rev. **D52** (1995) 5954.
- [25] N.E. Ligterink and B.L.G. Bakker, Phys. Rev. **D52** (1995) 5917.
- [26] S. Weinberg, Phys. Rev. **150** (1966) 1313.

- [27] E.E. Salpeter and H.A. Bethe, Phys. Rev. **84** (1951) 1232.
- [28] N. Nakanishi, Prog. Theor. Phys. Suppl. **43** (1969) 1; **95** (1988) 1.
- [29] C. Dietmayer et al., Z. Phys. **A334** (1989) 215,
- [30] Th. Heinzl, St. Krusche and E. Werner, Z. Phys. **A334** (1989) 443.
- [31] K.G. Wilson et al., Phys. Rev. **D49** (1994) 6720;
- [32] M.G. Fuda, Phys. Rev. **D41** (1990) 534; **D42** (1990) 2898; **D44** (1991) 1880; Ann. Phys. (N.Y.) **197** (1990) 265; **231** (1994) 1; Nucl. Phys. **A543** (1992) 111c.
- [33] J.M. Namyslowski, Phys. Rev. **D18** (1978) 3676.
- [34] P. Danielewicz and J.M. Namyslowski, Phys. Lett. **B81** (1979) 110.
- [35] P.M. Fishbane and J.M. Namyslowski, Phys. Rev. **D21** (1980) 2406.
- [36] J.M. Namyslowski and H.J. Weber, Z. Phys. **A295** (1980) 219.
- [37] M. Sawicki, Phys. Rev. **D32** (1985) 2666; **D33** (1986) 1103.
- [38] V.A. Karmanov, Nucl. Phys. **B166** (1980) 378.
- [39] G. Feldman, T. Fulton and J. Townsend, Phys. Rev. **D7** (1973) 1814.
- [40] A.N. Mitra, hep-ph/9812404; Phys.Lett.**B463**, 293 (1999).
- [41] A.N. Mitra and S. Bhatnagar, Int. J. Mod. Phys. **A7** (1992) 121.
- [42] A.N. Mitra et al., Phys. Rev. Lett., **59** (1987) 2408; Phys. Rev. **D38** (1988) 1454.
- [43] S. Chakrabarty et al., Prog. Part. Nucl. Phys. **22** (1989) 143.
- [44] A.N. Mitra, hep-th/9803062; Intl.J.Mod.Phys.**A14**, 4781 (1999).
- [45] V.R. Garsevanishvili, A.N. Kvinikhidze, V.A. Matveev, A.N. Tavkhelidze and R.M. Faustov, Teor. Mat. Fiz. **23** (1975) 310.
- [46] V.R. Garsevanishvili and V.A. Matveev, Teor. Mat. Fiz. **24** (1975) 3.
- [47] G.C. Wick Phys. Rev. **96** (1954) 1124;  
R.E. Cutkosky, Phys. Rev. **96** (1954) 1135.
- [48] S.J. Brodsky, C.-R. Ji and M. Sawicki, Phys. Rev. **D32** (1985) 1530.
- [49] J. Carbonell and V.A. Karmanov, Nucl. Phys. **A581** (1995) 625.
- [50] J. Carbonell and V.A. Karmanov, hep-ph/9812404; to appear in Euro. J. Phys. **A**.
- [51] V.A. Karmanov, Nucl. Phys., **A644** (1998) 165.
- [52] V.M. Kolybasov, Nucl. Phys. **68** (1965) 8.
- [53] Z. Dziembowsky, Phys. Rev. **D37** (1988) 768.
- [54] H.J. Melosh, Phys. Rev. **D9** (1974) 1095.
- [55] G.F. Gunion, S.J. Brodsky and R. Blankenbecler, Phys. Lett. **B39** (1972) 649; Phys. Rev. **D8** (1973) 287.
- [56] V. Glaser and B. Jaksic, Nuovo Cimento, **5** (1957) 1197.
- [57] V.A. Karmanov and A.V. Smirnov, Nucl. Phys. **A546** (1992) 691.

- [58] V.A. Karmanov and A.V. Smirnov, Nucl. Phys. **A575** (1994) 520.
- [59] V.A. Karmanov and J.-F. Mathiot, Nucl. Phys. **A602** (1996) 388.
- [60] M. Mangin-Brinet and J. Carbonell, "*Solution numerique du modele de Wick-Cutkosky dans le cadre de la Light Front Dynamics*", Rapport de Stage ISN/ECP (1997).
- [61] T. Frederico, private communication.
- [62] N.C.J. Schoonderwoerd and B.L.G. Bakker, Few-Body Syst. Suppl. **10** (1999) 119.
- [63] N.C.J. Schoonderwoerd, B.L.G. Bakker and V.A. Karmanov, Phys.Rev. **C58** (1998) 3093-3108
- [64] V.A. Karmanov, ZhETF, **83** (1982) 3 [JETP, **56** (1982) 1].
- [65] *Quantum Electrodynamics, selected papers*, Dover Publ. Inc., New York, 1958, ed. J. Schwinger.

# 31. 3D-4D Interlinkage Of B-S Amplitudes: Unified View Of $Q\bar{Q}$ And $QQQ$ Dynamics

A.N.Mitra \*

244 Tagore Park, Delhi-110009, India

## Abstract

This article has a 3-fold objective: i) to provide a panoramic view of several types of 3D vs 4D approaches in Field Theory (Tamm-Dancoff, Bethe Salpeter Equation (BSE), Quasi-potentials, Light-Front Dynamics, etc) for strong interaction dynamics; ii) to focus on the role of the Markov-Yukawa Transversality Principle (MYTP) as a novel paradigm for an exact 3D-4D interlinkage between the corresponding BSE amplitudes; iii) Stress on a closely parallel treatment of  $q\bar{q}$  and  $qqq$  BSE's stemming from a common 4-fermion Lagrangian mediated by gluon (vector)-like exchange. The two-way interlinkage offered by MYTP between the 3D and 4D BSE forms via a Lorentz-covariant 3D support to the BS kernel, gives it a unique status which distinguishes it from most other 3D approaches to strong interaction dynamics, which give at most a one-way connection. Two specific types of MYTP which provide 3D support to the BSE kernel, are considered: a) Covariant Instantaneity Ansatz (CIA); b) Covariant LF/NP ansatz (Cov.LF). Both lead to formally identical 3D BSE reductions (thus ensuring common spectral predictions), but their 4D manifestations differ sharply: Under CIA, the 4D loop integrals suffer from Lorentz mismatch of the vertex functions, leading to ill-defined time-like momentum integrals, but Cov LF is free from this disease. Some practical uses of MYTP as a basis for evaluating various types of 4D loop integrals are outlined.

PACS: 11.10 st ; 12.38 Lg ; 13.40.Fn

Keywords: Tamm-Dancoff, Bethe-Salpeter, Quasi-potentials, Light-front (LF), Markov-Yukawa, 3D-4D Interlinkage, CIA, Cov-LF, Spectroscopy.

## 1 Introduction: Effective BSE-SDE Framework

Ever since the success of the Tomonaga-Schwinger-Feynman-Dyson formalism in QED [1], corresponding field-theoretic formulations have been in the forefront of strong interaction dynamics since the early fifties, the main strategy being to device various 'closed' form approaches which are represented as appropriate 'integral equations'. One of the earliest efforts in this direction was the Tamm-Dancoff formalism [2] which showed a great intuitive appeal. In this method, the state vector of the system under consideration is Fock-expanded in terms of a complete set of eigenfunctions of the free field Hamiltonian, so that the expansion coefficients are the successive "amplitudes for finding present in the field a definite number of bare particles with definite spins and momenta...." [3]. This method was first systematically applied by Dyson (+ Cornell collaborators) in the early fifties, to the meson-nucleon scattering problem, for a dynamical understanding of the 'Delta' and other low-energy resonances; (see ref. [3]). It leads to 3D integral equations connecting amplitudes for successively higher numbers of meson (and/or nucleon-pair) quanta, much as the familiar 4D Schwinger-Dyson equations of QED connect (via Ward identities) vertex amplitudes of successively higher orders [4].

---

\*Email: (1)ganmitra@nde.vsnl.net.in; (2)anmitra@csec.ernet.in

### 1.1 3D Reduction of BSE: Quasipotentials, Etc

The 3D Tamm-Dancoff equation (TDE) [3] and the 4D Schwinger-Dyson equation (SDE) [4] have been the source of much wisdom underlying the formulation of many approaches to strong interaction dynamics. To these one should add the Bethe-Salpeter equation (BSE) [5], which is an approximation to SDE for the dynamics of a 4D two-particle amplitude, characterized by an effective (gluon-exchange-like) pairwise interaction, on the lines of a "Bethe Second Theory" of the Fifties for the effective N-N interaction, but now adapted to the quark level. Although not a fundamental (first principle) approach, such as the Chiral Schwinger Theory (CST) of  $(1+1)D$ , [see the Article by P.P.Srivastava in this Book], it has attracted more attention in the contemporary literature (as the 4D counterpart of the Schroedinger equation) than any other comparable approach. Perhaps it is a fair assessment that there is a degree of complementarity between first principle (emphasizing theoretical foundation) approaches like CST and second principle (stressing applicational aspects) approaches like BSE, in the sense that the lacunae of one are made up by the other, so that both together hold the key to a resolution of the strong coupling problem. In this Article we are concerned with the latter only.

A major bottleneck for the BSE approach has been its resistance to a probability interpretation, due to its 4D content. This has led to many attempts at its 3D reductions [6-9]: Instantaneous approximation [6]; Quasi-potential approaches [7,8]; variants of on-shellness of the associated propagators [9]. In [7,9], the starting BSE is 4D in all details, *including its kernel*, but the associated propagators are manipulated in various ways to reduce the 4D BSE to a 3D form as a fresh starting point of the dynamics; in [8], the old-fashioned 3D perturbation theory is reformulated covariantly to give a 3D quasipotential equation. These methods are briefly sketched in Sect 2.

At the 4D level, the BSE [5] is still the most widely used form of 2-particle dynamics, though the problems of probabilistic interpretation were the chief reason for these 3D formulations. Nevertheless the regular 4D equations of the full-fledged SDE-BSE types have been widely employed [10], as prototypes of strong interaction dynamics, addressing issues of gauge and chiral symmetries, as well as dynamical breaking of chiral symmetry via an NJL-type mechanism [11]. In such full-fledged field-theoretic approaches, the NJL-mechanism of contact interaction must of course be replaced by space-time extended interaction, and the dynamical breaking of chiral symmetry ( $DB\chi S$ ) corresponds to the use of SDE for the self-energy operator [12]. As a general remark, while for conceptual issues impinging on formulational self-consistency, there is little alternative to the full-fledged 4D equations, their applications to physical systems must recognize some ground realities. For example, the mass spectra of hadrons (which are revealed in Nature as  $O(3)$ -like [13]), suggest that the role of the time-like dimension (although an integral part of the dynamics) is not on the same footing as that of the space-like dimensions, so that a naive expectation of  $O(4)$ -like spectra [14] may be quite misleading. Indeed this issue is quite central to the very theme of this article, viz., 3D-4D interlinkage of BS-amplitudes, and will claim attention throughout.

An alternative form of 2-particle dynamics (which also contributes to reducing the effective degrees of freedom from 4D to 3D) is that of Dirac constrained Hamiltonian formalism [15], developed by Komar and others [16]. The logic of this approach is that constraints  $H_i$  have a twin role, viz., they not only 'constrain the motion' in phase space, but also generate it in their 'Hamiltonian' capacity. These 'constraints' must be mutually compatible in the sense  $[H_i, H_j] = 0$ . Such compatibility relations restrict the dependence of the interaction on the relative time  $t$ , and require a 'reciprocity relation' between the constituent potentials, something akin to Newton's III Law. Such descriptions are valid for both two-boson and two-fermion systems [16], in the sense of coupled Klein-Gordon and Dirac equations respectively. As this formalism is reviewed in detail in this Book by Lusanna [17a], it will not be discussed further here. In the same spirit, more fundamental approaches like the Chiral Schwinger Theory (CST) in  $(1+1)D$ , may be found in [17b].

### 1.2 Light Front(LF)/ Null Plane(NP) Dynamics

A powerful form of 3D dynamics came into prominence after Weinberg's discovery that the dynamics of the infinite momentum frame [18] serves as a cure for many ills in the theory of current

algebras, by greatly simplifying the rules of calculation of Feynman diagrams of old fashioned perturbation theory. In the present context of strong interaction dynamics, the great virtue of Weinberg's infinite momentum method [18] lies in the simplicity and transparency of the integral equations for multiparticle potential scattering problems [19]. Indeed, the structure of the 3-momenta  $(p_\perp, p_\parallel)$  appearing in this formalism is but a paraphrase for the standard null-plane variables first introduced by Dirac to project his theme [20] that a relativistically invariant Hamiltonian theory can be based on 3 different classes of initial surfaces (space-like, time-like, and null-like). The structure of such a Hamiltonian theory is strongly dependent on these respective surface forms whose "stability groups" (i.e., those generators of the Poincare group that leave the initial surface *invariant*), are 6, 6, 7 respectively, thus giving the 'highest score' (7) to the null-plane dynamics ( $x_0 = x_3$ ) whose 'kinematical' generators form a closed algebra, and include among others the quantity  $P_+ = P_0 + P_3$  (which plays the role of the 'mass' term  $\eta$  in the Weinberg notation [18]). On the other hand, the dual generator  $P_- = P_0 - P_3$  is the 'Hamiltonian' of the theory.

Leutwyler and Stern [21] gave a covariant 3D formulation of the Dirac theme [20] in terms of null-plane variables. A more explicit covariant formulation in the null-plane language was given by Karmanov [22a] using diagrammatic techniques with on-shell propagators and spurions, on the lines of the Kadychevsky approach [8], which has been recently reviewed by Carbonell et al [22b], (referred to as KK). All these methods, including the Wilson group's [23], give rise to 3D integral equations for a (strongly interacting) two-body system, bearing strong resemblance to the other 3D BSE forms [6-9] above. Again, there is no getting back to the original 4D BSE form, the nearest connection being a one-way reduction from 4D BSE to 3D on the covariant null-plane [22b].

### 1.3 Markov-Yukawa Transversality: 3D-4D Interlinkage

Finally we come to a rather novel approach of more recent origin [24-25], based on the Markov-Yukawa Transversality Principle (MYTP) [26]. To motivate this approach, it is necessary to go back in time to Yukawa's non-local field theory [26b] according to which the field variable is a function of both  $x$  and  $p$ , unlike in local field theory in which the field variable is only a (local) function of  $x$ . Although unacceptable for an elementary particle/field, the non-local field theory is ideally suited to a *composite* particle, whose extended structure effectively provides for a momentum dependence in the direction of the total 4-momentum  $P_\mu$ . Indeed the Yukawa theory [26b] was in a way the forerunner of a later theory of bi-local fields  $\mathcal{M}(x, y)$  [27] for the formulation of the Effective Action for a 2-body dynamical system [27]. This approach was employed by the Prevushin group [25] in their formulation of the relativistic Coulomb problem in the Salpeter approximation [6b] in a covariant form, with the choice of the preferred direction governed by the 4-momentum  $P_\mu$  of the composite as the canonical conjugate to its c.m. position  $X = (x + y)/2$ :  $P = -i\partial_X$ . More specifically the MYTP [26] is expressed by the condition [25]:

$$z_\mu \frac{\partial}{\partial X_\mu} \mathcal{M}(z, X) = 0; \quad z = x - y \quad (1)$$

where the direction  $P_\mu$  guarantees an irreducible representation of the Poincare' group for the bilocal field  $\mathcal{M}$  [26c]. This condition in turn is equivalent to a covariant 3D support to the input 4-quark Lagrangian, whence follow the SDE and BSE as equations of motion with 3D support to the effective BSE kernel under covariant instantaneity.

Alternatively, the 3D support ansatz may be directly postulated at the outset for the pairwise BSE kernel  $K$  [24] by demanding that it be a function of only  $\hat{q}_\mu = q - q.PP_\mu/P^2$ , which implies that  $\hat{q}.P \equiv 0$ . In this approach, the propagators are left untouched in their full 4D forms. This is somewhat complementary to the approaches [6-9] (propagators manipulated but kernel left untouched), so that the resulting equations [24-25] look rather unfamiliar vis-a-vis 3D BSE's [6-9], but it has the advantage of allowing a *simultaneous* use of both 3D and 4D BSE forms via their interlinkage. Indeed what distinguishes the Covariant Instantaneity Ansatz (CIA) [24] from the more familiar 3D reductions of the BSE [6-9] is its capacity for a 2-way linkage: an exact 3D BSE reduction, and an equally exact reconstruction of the original 4D BSE form without extra

charge [24]: the former to access the observed  $O(3)$ -like spectra [13], and the latter to give transition amplitudes as 4D quark loop integrals [24]. (In the approach of the Pervushin Group [25], however, the built-in 3D-4D interconnection which follows from MYTP [26], apparently remained unnoticed in their final equation). In contrast the more familiar methods of 3D give at most a one-way connection, viz., a  $4D \rightarrow 3D$  reduction [6-9], but not vice versa.

At this point it is perhaps worth noting that even the Salpeter equation [6b] had (in principle) the ingredients for a reconstruction of the 4D BS amplitude  $\Psi$  in terms of 3D ingredients, provided the ‘instant’ form, (see eq.(2.1) in Sect 2), of the interaction kernel had been employed on the RHS of the 4D BSE form, and simultaneously the 4D BS amplitude  $\Psi$  on the RHS had been eliminated in favour of the 3D BS amplitude  $\psi$ , exactly as was done under CIA [24]. This would have amounted to using the Transversality Principle [26] (albeit non-covariantly), but this feature had apparently remained unnoticed by subsequent workers who continued to employ the Salpeter equation [6b] in its 3D form only.

## 1.4 QCD-motivated Effective Lagrangians

The Transversality Principle (MYTP) [26] underlying the 3D-4D interconnection [24], termed 3D-4D-BSE in the following, of course needs supplementing by physical ingredients to govern the structure of the BSE kernel, much as a Hamiltonian needs a properly defined ‘potential’. However its canvas is broad enough to accommodate a wide variety of kernels which must in turn be governed by independent physical principles. In this respect, short of a full-fledged QCD Lagrangian approach, the orthodox view (which we adopt) is to stick to an effective 4-fermion Lagrangian as a starting point of the dynamics, from which the successive equations of motion (SDE, BSE, etc) follow in the standard manner. [As already noted at the outset, this is in keeping with the Bethe Second Principle Theory for effective  $N - N$  potentials as an input for the physics of the nuclear many-body problem].

In particular, a basic proximity to QCD is ensured through a vector- type interaction [12], which while maintaining the correct one-gluon-exchange structure in the perturbative region, may be fine-tuned to give any desired structure to the intermediate gluon propagator in the infrared domain as well. Although empirical, it captures a good deal of physics in the non-perturbative domain while retaining a broad QCD orientation, albeit short of a full-fledged QCD formulation. More importantly, the non-trivial solution of the SDE corresponding to this generalized gluon propagator [12] gives rise to a dynamical mass function  $m(p)$  [12] as a result of  $DB\chi S$ , w.r.t. an input Lagrangian whose chiral invariance stems from a vector-type 4-fermion interaction between almost massless  $u - d$  quarks. These considerations form the standard basis for a Lagrangian-based BSE-SDE framework [10] for Dynamical Breaking of Chiral Symmetry ( $DB\chi S$ ) [11], for a space-time extended 4-quark Lagrangian mediated by vector exchange [12]. This generates a mass- function  $m(p)$  via Schwinger-Dyson equation (SDE), which accounts for the bulk of the constituent mass of  $ud$  quarks. The same BSE-SDE formalism [12,10] can be simply adapted [28] to the MYTP [26]-based 3D-4D-BSE formalism [24] which reproduces 3D spectra of both hadron types [29] under a common parametrization [28] for the gluon propagator, with a self-consistent SDE determination [28] of the constituent mass; see Sect 3.

A BSE-SDE formulation [10] on QCD lines represents a 4D field- theoretic generalization of ‘potential models’ [30], wherein the generalized 4-fermion kernel [12] represents the non-perturbative gluon propagator, which can be easily adapted [28] to MYTP [26]). The 4D feature of BSE-SDE gives this framework a ready access to high energy amplitudes, while its ‘off-shell’ features give it a natural access to hadronic spectra [13]. It has thus an interpolating role between (low energy) quarkonia models [30], and (high energy) QCD-SR [31] techniques whose domains are largely complementary; details may be found in a recent review [32].

## 1.5 MYTP: Cov Instantaneity vs Cov LF/NP

While MYTP [26] ensures 3D-4D interconnection [24] under covariant instantaneity ansatz (CIA) in the composite’s rest frame [24], its main disadvantage lies in the ill-defined nature of 4D loop



integrals which acquire time-like momentum components in the exponential/gaussian factors associated with the different vertex functions, due to a ‘Lorentz-mismatch’ among the rest-frames of the participating hadronic composites. This problem is especially serious for triangle loops and above, such as the pion form factor, while 2-quark loops [33] just escape this pathology. This problem was not explicitly encountered in the light-front (LF/NP) ansatz [34] in an earlier study of 4D triangle loop integrals, but this approach was criticized [35] on grounds of non-covariance. The CIA approach [24] which made use of MYTP [26], was an attempt to rectify the Lorentz covariance defect, but the presence of time-like components in the gaussian factors inside triangle loop integrals, e.g., in the pion form factor [36], impeded further progress.

In an attempt to remedy this situation, a generalization of MYTP [26] was proposed recently [37] to ensure formal covariance without having to encounter time-like components in the gaussian wave functions appearing inside the 4D loop integrals. The desired generalization was achieved by extending the Transversality Principle [26] from the covariant rest frame of the (hadron) composite [24], to a *covariantly defined* light-front [37] (Cov LF). It was found that while preserving the 3D-4D BSE interconnection, the resulting 3D equation under Cov LF [37] turns out to be formally identical to the old-fashioned null-plane formalism [34,38], so that the latter enjoys *ipso facto* covariance (despite its ‘looks’). This ‘covariant’ LF/NP method [37] stands fairly direct comparison with other covariant LF approaches [22-23].

## 1.6 Scope of the Article

This article has a 3-fold objective: A) a bird’s eye view of some principal 3D vs 4D dynamical methods for the strong interaction problem that have been proposed over the last half century; B) Putting in perspective a novel property of the Markov-Yukawa Transversality Principle (MYTP), viz., a 2-way 3D-4D interconnection in the BS dynamics of 2- and 3-quark hadrons; C) Stressing a close parallelism between  $q\bar{q}$  and  $qqq$  BSE’s which stem from a common 4-fermion Lagrangian mediated by gluon (vector)-like exchange. Especially noteworthy is the capacity of MYTP [26] to achieve a 3D-3D interlinkage, a property which has remained obscured from view in the contemporary literature, vis-a-vis more familiar approaches to BSE and allied forms of dynamics [6-10, 18-23] which are either 3D or 4D in content, but have no provision for any interlinkage between these two dimensions. While the details of individual MYTP applications (several of which are dealt with in [32]), are not a part of this Article, the practical uses of MYTP in the strong interaction dynamics regime will nevertheless be a focus of attention by virtue of its distinct advantage in the evaluation of 4D loop integrals with arbitrary vertex functions, while providing easy access to the spectroscopy sector. To that end, an outline of the dynamical structure of some principal 3D methods, [7-9], [18-23], is provided in Sect.2 as a background for comparison on 4D loop integral techniques, while on the issue of hadron spectra, which are basically  $O(3)$ -like [13], a comparison between non-MYTP [7-9] and MYTP [24-26] forms of dynamics does not bring out new physics.

For a better understanding of the working of MYTP, it will be necessary to present two types in parallel for comparison, viz., Covariant Instantaneity or CIA [24-25], and Covariant Light-front (Cov LF) [37], which demand that the BSE kernel  $K$  for pairwise interaction be a function of relative momentum  $q$  *transverse* to the composite 4-momentum in the first case [24], and to the Covariant Null-plane in the second [37]. It will be shown that both types lead to identical 3D BSE reductions (so that their spectral predictions are formally the same), but their reconstructed 4D vertex functions reveal profound differences in structure: The Lorentz mismatch of individual wave functions that characterizes the CIA form [24], leading to complex amplitudes [34], disappears in the alternative Cov LF approach [37], but in general such integrals are dependent on the light-front orientation  $n_\mu$ , as in other covariant approaches [22-23]. To eliminate such terms, a simple prescription of ‘Lorentz completion’ seems to suffice to produce an explicitly Lorentz invariant quantity such as was shown for the pion form factor [37]; (alternative prescriptions exist in other LF/NP formulations [22b]). For a historical perspective, it is useful to recall that in the old-fashioned NPA approach too [38], a very similar result had been found for various types of triangle loop amplitudes [36], despite a lack of manifest covariance [35] in that approach, but now MYTP [26] on the covariant null-plane [37] fills this formal gap.

## 1.7 Outline of Contents

Sect.2 briefly outlines some historical approaches to an effective 3D form of strong interaction dynamics: Levy-Salpeter [6]; Logunov- Tavkhelidze [7a]; Blankenbecler-Sugar [9]; Todorov [7d] ; Weinberg [18]; Feynman et al [39]. Sect.3 provides the theoretical framework with a short derivation of the BSE-SDE from an input chirally invariant Lagrangian, incorporating the original CIA form [25,24] of MYTP [26], on the lines of ref.[25] in terms of bilocal fields [27]. It also includes a derivation [28] of the dynamical mass function  $m(p)$  for an understanding of the constituent mass via Politzer additivity [40]. Sect.4 collects some basic results on the null-plane formalism due to Leutwyler-Stern [21], especially the role of the ‘Angular Condition’ in ensuring a formal  $O(3)$ -like invariance. From Sect.5 onwards, the focus is on MYTP [26]-orientation for bringing out its unique property which distinguishes it from most other approaches, viz., the *3D-4D interlinkage* of BS amplitudes.

Sect.5 gives a comparative view of the working of MYTP on the BSE forms in CIA [24] versus Cov LF [37], and outlines the derivation of the 3D BSE, as well as an explicit reconstruction of the 4D BS wave function in terms of 3D ingredients, with 3-momentum  $\hat{q} = (q_\perp, q_3)$ , where the third component emerges as a  $P$ -dependent one, suitably adapted to the CIA [24] or Cov LF [37] respectively. Sec.6 gives a corresponding derivation for the 3D-4D interconnection for a  $qqq$  BSE structure under CIA conditions. Sec.7 illustrates, through the calculation of triangle-loop integrals, the relative advantage of Cov LF [37] over the CIA [24] version of MYTP [26], in producing a well-defined structure for the pion e.m. form factor in a fully gauge invariant manner, and illustrating in the process the method of ‘Lorentz-completion’ for explicit Lorentz invariance, with the expected  $k^{-2}$  behaviour at high  $k^2$ . MYTP also gives a more general structure of triangle loop integrals for three-hadron form factors. Sect.8 summarises our conclusions.

## 2 Quasipotentials And Other 3D Dynamical Equations

The reduction of the 4D BSE for an  $N-N$  pair to the 3D level in the Instantaneous Approximation was first investigated in the non-adiabatic domain of pseudoscalar meson theory (effect of pair-creations included) by Levy [6a], who showed that this 3D BSE form is entirely equivalent to the corresponding Tamm-Dancoff equations [2] in the same (non-adiabatic) limit.

On the other hand, Salpeter [6b] employed the adiabatic approximation (no pair creation effects) to give a systematic 3D reduction of the fermionic BSE, using projection operators for large and small components. The adiabatic approximation amounts to replacing the propagator  $\Delta_F(x-x')$  for the exchanged meson by

$$\Delta_F(x-x') \Rightarrow \delta(x_0-x'_0) \int_{-\inf}^{\inf} \Delta_F(\mathbf{x}-\mathbf{x}', x_0-x'_0) d(x_0-x'_0) \quad (2.1)$$

and simply gives the Yukawa potential between two particles. Similarly, in the Instantaneous (adiabatic) Approximation, IA for short, the 4D wave function  $\Psi(x) = \Psi(\mathbf{x}, t)$  for relative motion of two particles becomes simply  $\Psi(\mathbf{x}, 0)$ . In the momentum representation, these statements read respectively as

$$\Delta_F(k) \Rightarrow \Delta_F(\mathbf{k}); \quad \psi(\mathbf{q}) = \int dq_0 \Psi(\mathbf{q}, q_0) \quad (2.2)$$

The Salpeter 3D BSE in the IA for a relativistic hydrogen-atom is [6b]:

$$(E - H_1(\mathbf{q}) - H_2(\mathbf{q}))\chi(\mathbf{q}) = \int d^3k \frac{e^2}{2\pi^2(\mathbf{k}-\mathbf{q})^2} [\Lambda_{1+}\Lambda_{2+} - \Lambda_{1-}\Lambda_{2-}] \chi(\mathbf{k}) \quad (2.3)$$

where the 3D wave function  $\chi(\mathbf{q})$  is related to the corresponding 4D quantity by an equation of the form (2.2), and the symbols  $\Lambda_\pm$  are energy projection operators for the large/small components, etc.

## 2.1 Logunov-Tavkhelidze Quasipotentials

A different form of 3D reduction of the 4D BSE was proposed by Logunov- Tavkhelidze [7a] in the language of Green's functions (G-fns) for 2-particle scattering whose momentum representation may be written as  $G(p_1 p_2; p'_1 p'_2)$  (with indicated 4-momenta before and after), which satisfies a 4D BSE [7a]:

$$(2\pi)^8 \Delta(p_1) \Delta(p_2) G(p_1 p_2; p'_1 p'_2) = \delta(p_1 - p'_1) \delta(p_2 - p'_2) + \int d p''_1 d p''_2 K(p_1 p_2; p''_1 p''_2) G(p''_1 p''_2; p'_1 p'_2) \quad (2.4)$$

where  $\Delta(p_i) = p_i^2 + m_i^2$ , etc. Expressing this equation in c.m. ( $P$ ) and relative ( $q$ ) 4-momenta, and taking out the  $\delta$ -fns due to the c.m. motion, this equation simplifies

$$(2\pi)^4 \Delta(p_1) \Delta(p_2) G(q, q'; P) = \delta(q - q') + \int d q'' K(q, q'') G(q'', q; P) \quad (2.5)$$

Next, they defined the 3D G-fn for the relative motion as a double integral w.r.t. the two time-like momenta:

$$\hat{G}(q, q'; P) = \int q_0 \int q'_0 G(q, q'; P) \quad (2.6)$$

Now in operator notation, the 4D BSE (2.5) may be written as  $G = G_0 + G_0 K G$ , from which the kernel  $K$  has the formal representation  $K = G_0^{-1} - G^{-1}$ . The L-T trick [7a] now consists in using the double integrals on the time-like momenta as in eq.(2.6) to formally define the 3D quasipotential  $\hat{K}$  as

$$\hat{K} \equiv \hat{G}_0^{-1} - \hat{G}^{-1} \quad (2.7)$$

which can be expanded perturbatively in the symbolic form [7a]

$$\hat{K} = \hat{G}_0^{-1} G_0 \hat{K} G_0 \hat{G}_0^{-1} - \hat{G}_0^{-1} G_0 K \hat{G}_0 K G_0 \hat{G}_0^{-1} - \dots \quad (2.8)$$

to any desired order of accuracy; [note that the inverse G-fns are just the self-energy operators]. If  $V(\hat{q}, \hat{q}'; E)$  is the quasi-potential to a given order of accuracy, then, the BSE satisfied by the 3D BS wave function  $\psi(\hat{q})$  is of the form [7a]:

$$(E^2 - \hat{q}^2 - m^2) \psi(\hat{q}) = \int d^3 \hat{q}' V(\hat{q}, \hat{q}'; E) \psi(\hat{q}') \quad (2.9)$$

where the 'denominator function on the LHS arises from integrating  $G_0 = \Delta(p_1)^{-1} \Delta(p_2)^{-1}$  w.r.t.  $q_0$  and rearranging.

### 2.1.1 Narrow resonances in charged particle systems

Within the last decade, the L-T theory [7a] has witnessed some interesting applications [41] to the understanding of 'new' narrow  $e^+e^-$  resonances observed in heavy ion collisions [42]. To that end, the authors [41] have employed an equation of the form (2.9) which reads for this system as [41]:

$$2\omega(M - 2\omega)\phi(\mathbf{p}) = \frac{(2me)^2}{(2\pi)^3} \int \frac{d^3 \mathbf{p}' \phi(\mathbf{p}')}{2\omega' q(M - \omega - \omega' - q + i0)} \quad (2.10)$$

where  $\omega = \sqrt{m^2 + \mathbf{p}^2}$ , and  $q = |\mathbf{p} - \mathbf{p}'|$ . The results indicate a possible interpretation of the observed peaks [42] as new quasi-stationary levels arising from the solution of the quasi-potential equation. More interestingly, they also suggest a close relationship of the observed states [38] with the von Neumann-Wigner [43] levels embedded in the continuum.

## 2.2 Blankenbecler-Sugar Equation

Another type of quasipotential was proposed by Blankenbecler-Sugar [9], as follows. The 2-particle scattering amplitude  $T(q, q')$  due to a 4D potential  $V(q, q')$  in the ladder approximation satisfies the BSE [9]:

$$T(q, q') = -i(2\pi)^{-4} \int d^4 q'' V(q, q'') [m^2 + (P/2 + q'')^2]^{-1} [m^2 + (P/2 - q'')^2]^{-1} T(q'', q') \quad (2.11)$$

where the 2-particle 'free' G-fn is exhibited as the product of the two propagators inside the integral on the RHS. To express this equation in 3D form, the B-S [9] trick consists first in putting  $q'$  on the energy shell, which means that  $q'_0 = 0$ , and  $q'^2 = s/4 - m^2$ , where  $s = -P^2$  is the square of the c.m. energy. Next, the on-shell part  $E_2$  of the free 2-particle G-fn is obtained by taking only the  $\delta$ -fn parts of the two propagators which gives rise only to two-particle cuts in the physical region

$$E_2(q'') = 2\pi \int ds' (s' - s)^{-1} \delta[m^2 + (P'/2 + q'')^2] \delta[m^2 + (P'/2 - q'')^2] \quad (2.12)$$

where  $s' = -P'^2$ , and  $P'$  has only a fourth component. This works out as

$$E_2(q'') = \frac{1}{2} \pi \delta(q''_0) [\sqrt{(q''^2 + m^2)} (q''^2 - q^2)]^{-1} \quad (2.13)$$

The balance  $R_2$  of the free G-fn is not singular along the positive cut of the  $s$ -variable. If it is neglected in the first approximation, and only  $E_2$  from (2.13) is substituted in (2.11), then after a trivial integration over  $q''_0$ , the resultant 3D equation has the form

$$T(q, q') = V(q, q') + \frac{1}{4} \int \frac{d^3 q''}{(2\pi)^3} \frac{V(q, q'') T(q'', q')}{\sqrt{m^2 + q''^2} (q''^2 - q'^2)} \quad (2.14)$$

A comparison of (2.9) and (2.14) shows that although both equations are formally 3D in looks, there is a vast difference in their contents: The L-T [7a] form (2.9) involves only 3-momenta  $\hat{q} \equiv \mathbf{q}$ , since the Hilbert space has been 'truncated' by integrating out over their fourth components. The B-S [9] form (2.14) on the other hand has 4-momenta formally throughout (no truncation of Hilbert space), except that they are on their mass shells! Thus formal covariance is violated in both equations, although in different ways.

## 2.3 Kadychevsky-Todorov Equation

Still another form of 3D (Lippmann-Schwinger) equation was given by Todorov [7d], following the Covariant method of Kadychevsky [8]. In the Todorov approach [7d], the potential  $V_w$  is defined as an infinite power series in the coupling constant which fits the perturbative expansion of the scattering amplitude  $T_w$  for two particles of masses  $m_1, m_2$  and 4-momenta  $p_1, p_2$  and  $q_1, q_2$  before and after respectively. The quantity  $T_w$  in the off-shell regime satisfies the L-S equation [7d]

$$T_w(\mathbf{p}, \mathbf{q}) + V_w(\mathbf{p}, \mathbf{q}) + \frac{1}{\pi^2 w} \int d^3 \mathbf{k} \frac{V_w(\mathbf{p}, \mathbf{q}) T_w(\mathbf{p}, \mathbf{q})}{\mathbf{k}^2 - b^2 - i\epsilon} \quad (2.15)$$

where the 3D quantities in the c.m. frame are defined as

$$\mathbf{p}_1 = -\mathbf{p}_2 = \mathbf{p}; \quad \mathbf{q}_1 = -\mathbf{q}_2 = \mathbf{q} \quad (2.16)$$

and on the energy shell, the corresponding time-like quantities are

$$p_{10} + p_{20} = w = q_{10} + q_{20}; \quad -p^2 = -q^2 = w^2; \quad 4w^2 b^2(w) = \lambda(w^2, m_1^2, m_2^2) \quad (2.17)$$

This equation too has strong resemblance to the L-T equation [7a]. The corresponding equation for the bound state wave function  $\phi(\mathbf{p})$  is

$$\pi^2 w (\mathbf{k}^2 - b^2(w)) \phi(\mathbf{p}) = - \int d^3 \mathbf{k} V(\mathbf{p}, \mathbf{k}) \phi(\mathbf{k}) \quad (2.18)$$

Both B-S [9] and Todorov[7d] equations have been extensively employed in the literature.

## 2.4 Infinite-Momentum Frame: Weinberg Equation

Weinberg [18] observed some remarkable simplifications that occur when the results of old-fashioned perturbation theory are expressed in a reference frame in which the total 3-momentum  $\mathbf{P}$  is very large. In this limit, the 3-momentum  $\mathbf{p}_n$  of the  $n$ -th particle may be projected parallel and perpendicular to  $\mathbf{P}$ , and the results collected as follows:

$$\mathbf{p}_n = \eta_n \mathbf{P} + \mathbf{q}_n; \quad \mathbf{q}_n \cdot \mathbf{P} = 0; \quad \sum_n \eta_n = 1; \quad \sum_n \mathbf{q}_n = 0. \quad (2.19)$$

The quantity  $\eta_n > 0$  in this theory, plays the role of 'mass' of the  $n$ -th particle (in a 3D Schrodinger-type equation), and in the  $P \rightarrow \infty$  limit, the rules of calculation become very simple: all old-fashioned perturbative diagrams passing through negative energy intermediate states vanish, while for the contributing diagrams, the propagator for an intermediate state  $c$  in a transition from  $a$  to  $b$ , has the form  $2[s_a - s_c + i\epsilon]^{-1}$ , where  $s$  for any state is the usual total c.m. energy squared:

$$s = \sum_n [\mathbf{q}_n^2 + m_n^2]/\eta_n; \quad s_a = s_b = s_c, \text{ etc.} \quad (2.20)$$

Momentum conservation at each vertex is 3D in content:

$$(2\pi)^3 \delta(\Delta \sum \eta) \delta^2(\Delta \sum \mathbf{q}) \quad (2.21)$$

in accordance with the conservation of  $\eta$  and  $\mathbf{q}$ , eq.(2.15). The Weinberg counterpart of the L-T [7a], B-S [9] and Todorov [7d] equations (2.9), (2.14) and (2.18) respectively, is the integral equation [18]

$$\langle \mathbf{q}' \eta' | T | \mathbf{q} \eta \rangle = \langle \mathbf{q}' \eta' | V | \mathbf{q} \eta \rangle + \int d^2 q'' \int d\eta'' \frac{\langle \mathbf{q}' \eta' | V | \mathbf{q} \eta \rangle \langle \mathbf{q}' \eta' | V | \mathbf{q} \eta \rangle}{2(2\pi)^3 [s\eta''(1-\eta'') - \mathbf{q}''^2 - m^2 + i\epsilon]} \quad (2.22)$$

Although this equation is effectively 3D, and has considerable similarity to the corresponding equations of [7a,7d,9] above, there is a big difference, viz., the angular momentum is no longer a well-defined concept in this 3D description. This gap was bridged later by Leutwyler-Stern [21] by invoking the 'angular condition' [21], after it became clear that the Weinberg approach is equivalent to Dirac's [20] null-plane dynamics; see Sect.3 below.

## 2.5 The FKR Model For 2- And 3-Quark Dynamics

Before ending this Section, we draw attention for historical reasons, to a unique paper by Feynman and collaborators [39], FKR for short, which gave an integrated view of 2- and 3-quark hadron dynamics, and played a big role in shaping the direction of strong interaction physics to come. The importance of the FKR approach stems, among other things, from the fact that these authors were the first to show the way to a unified treatment of both 2- and 3-quark hadrons within a common dynamical framework, which was to serve as a model for the future. This paper effectively incorporated all the relevant aspects of quark dynamics that had been generated piecemeal in the Sixties, and had by and large come to be accepted, viz., the group structure  $SU(6) \times O(3)$ , the symmetrical quark model, and harmonic oscillator classification of hadron states (based on their linear  $M^2-N$  plots) on the one hand [44], and the mechanism of single-quark transitions, quark recoil effects, etc, on the other [45].

The FKR model, which made essential use of harmonic confinement, sought to give a relativistic meaning to the internal motion of light quarks through the following definitions of 2- and 3-quark Hamiltonians [39, 38]:

$$-K_M = 2(p_1^2 + p_2^2) + \frac{1}{16} \Omega^2 (x_1 - x_2)^2 + \text{Const} \equiv P^2 + M_M^2; \quad (2.23)$$

$$-K_B = 3(p_1^2 + p_2^2 + p_3^2) + \frac{1}{36} \Omega^2 \sum_{123} (x_1 - x_2)^2 + \text{Const} \equiv (P^2 + M_B^2) \quad (2.24)$$

where  $x_{1\mu} = i\partial_{1\mu}$ ;  $p_1^2 = p_{1\mu}p_{1\mu}$ , etc. The quantity  $\Omega$  which is postulated to be the *same* for *both* systems, has the significance of the universal Regge slope ( $\approx 1\text{GeV}^2$ ) as observed [44] in their respective spectra; [Note the geometrical factors as coefficients in front of the respective kinetic and potential terms above]. The operators  $K_{M,B}^{-1}$  are the ‘free’ propagators (albeit with h.o. confinement) for the mesons and baryons, whose ‘poles’ correspond to the eigenvalues (spectra) of their squared masses. The presence of a perturbation  $\delta K$  can be simulated in a standard gauge-invariant manner, by the substitutions  $p \rightarrow p_\mu - eV_\mu$  or  $p_\mu \rightarrow p_\mu - g\gamma_5 A_\mu$  for vector and axial vector couplings respectively, after rewriting  $p_\mu^2$  as  $(\gamma \cdot p)^2$ , while the  $i \rightarrow j$  transition amplitudes are just  $\langle h_j | \delta K | h_i \rangle$ , by standard rules of quantum mechanics.

A major achievement of the FKR model was its success in giving two distinct types of unification, viz., a common framework for Spectroscopy and transition amplitudes; and ii) a unified dynamical treatment of  $q\bar{q}$  and  $qqq$  hadrons. Both these features represented landmarks in a dynamical understanding of the quark model, yet the FKR model failed on another count: the ‘wrong’ sign of the time-like momenta in the gaussian wave functions for the hadrons was a disease which pointed to an asymmetric role of time-like (1D) momenta vis-a-vis the space-like (3D) ones. Attempts to cure this disease by a Euclidean treatment (via Wick rotation) [46a] failed on the spectroscopy front [13] which reveal only  $O(3)$ -like spectra, while other non-covariant treatments [46b] were not very successful either. Nevertheless the lessons from the FKR model were significant pointers to the need to treat the 1D time-like and 3D space-like d.o.f.’s on different footings in a future quest for a covariant theory [24-26].

### 3 Self-Energy And Vertex Fns Under MYTP

As a first step towards introducing the MYTP [26] theme, we collect in this Section some essential machinery for the interconnection between self-energy and vertex functions via Schwinger-Dyson (SDE) and Bethe-Salpeter (BSE) equations, starting from a chirally invariant Lagrangian characterized by a vector-type interaction [12] as a prototype for a gluon-exchange propagator in the non-perturbative QCD regime [28]. To that end, we shall first outline the method of bilocal fields [27] to derive the equations of motion (SDE and BSE) from such a Lagrangian, following the Pervushin Group’s [25] bilocal field method, under MYTP [26] conditions of covariant instantaneity [24]. This will be followed by a general result connecting self-energy and pion-quark vertex functions in the chiral limit, i.e., when the pion mass vanishes. This result in turn paves the way to a derivation [28], under MYTP [26] conditions of Covariant Instantaneity [24], of the mass function  $m(p)$  whose low momentum limit is the main contributor to the constituent mass, via Politzer additivity [40].

#### 3.1 Method of Bilocal Fields for BSE-SDE

The effective action for a system of two interacting massless fermions constrained by MYTP [26] is given by [25]

$$W_{eff}[\psi, \bar{\psi}] = \int d^4x [\bar{\psi}(x)(i\gamma\partial - m_0)\psi(x) + \frac{1}{2} \int d^4y (\psi(y)\bar{\psi}(x))K(z^\perp, X)(\psi(x)\bar{\psi}(y))] \quad (3.1)$$

where  $z = x - y$ ;  $X = (x + y)/2$ .  $z^\perp$  is the component of  $z$  transverse to the  $P$ -direction. Now redefine the action (3.1) in terms of bilocal fields  $\mathcal{M}$  via the Legendre transformation [27] on the second term to give

$$-\frac{1}{2} \int d^4x d^4y \mathcal{M}(x, y) K^{-1}(z^\perp, X) \mathcal{M}(x, y) + \int d^4x d^4y (\psi(x)\bar{\psi}(y)) \mathcal{M}(x, y) \quad (3.2)$$

Then in an obvious short-hand notation [25b], (3.1) may be written as

$$W_{eff}[\mathcal{M}] = (\psi\bar{\psi}, (-G_0^{-1} + \mathcal{M}) - \frac{1}{2}(\mathcal{M}, K^{-1}\mathcal{M})) \quad (3.3)$$

where  $G_0$  is the inverse Dirac operator for the free fermion field. After quantization over  $N_c$  fermion fields and normal ordering, the action takes the form [25b]

$$W_{eff}[\mathcal{M}] = -\frac{1}{2}N_c(\mathcal{M}, K^{-1}\mathcal{M}) + iN_c \sum_{n=1}^{\infty} \frac{1}{n} \Phi^n \quad (3.4)$$

where  $\Phi = G_0\mathcal{M}$  is a matrix in  $(x, y)$  space, and its successive powers are defined in the standard matrix fashion. Now for the quantization of the action (3.4), its minimum is given by

$$N_c^{-1} \frac{\delta W_Q(\mathcal{M})}{\delta \mathcal{M}} \equiv -K^{-1}\mathcal{M} + \frac{1}{G_0^{-1} - \mathcal{M}} = 0. \quad (3.5)$$

The corresponding 'classical' (lowest order) solution for the bilocal field is  $\Sigma(x-y)$  which depends only on the difference  $x-y$  due to the translation invariance of the vacuum solutions. Next expand the action (3.4) around the point of minimum  $\mathcal{M} = \Sigma + \mathcal{M}'$ , and denote the small fluctuations  $\mathcal{M}'$  as a sum over the complete set of 'classical' solutions  $\Gamma$ . Then in the next order of extremum, we have:

$$\frac{\delta^2 W_Q(\Sigma + \mathcal{M}')}{\delta \mathcal{M}^2} \big|_{\mathcal{M}'=0} \Gamma = 0 \quad (3.6)$$

Eqs.(3.5-6) give respectively the SDE for  $\Sigma$  and BSE for  $\Gamma$ :

$$\begin{aligned} \Sigma(x-y) &= m_0 \delta^4(x-y) + iK(z^\perp, X)G_\Sigma(x-y); \\ \Gamma &= iK(z^\perp, X) \int d^4z_1 d^4z_2 G_\Sigma(x-z_1)\Gamma(z_1, z_2)G_\Sigma(z_2-y) \end{aligned} \quad (3.7)$$

which describe the spectrum of the fermions and composites respectively. In momentum space these equations for the mass operator and vertex function are

$$\Sigma(\hat{p}) = m_0 + i \int \frac{d^4q}{(2\pi)^4} V(\hat{p}-\hat{q})\gamma\hat{\eta}G_\Sigma(q)\gamma\hat{\eta}; \quad \eta_\mu = P_\mu/|P| \quad (3.8)$$

$$\Gamma(\hat{k}) = i \int \frac{d^4q}{(2\pi)^4} V(\hat{k}-\hat{q})\gamma\hat{\eta}[G_\Sigma(q+P/2)\Gamma(q^\perp)G_\Sigma(q-P/2)]\gamma\hat{\eta} \quad (3.9)$$

where  $G_\Sigma(q) = (\gamma\hat{q} - \Sigma(q^\perp))^{-1}$ ;  $V$  is the scalar part of the kernel  $K$  with 3D support;  $\hat{k}$  is the transverse part of  $k$  w.r.t. the direction  $\eta_\mu$  of the total 4-momentum  $P_\mu$ .

### 3.2 Self-Energy vs Vertex Fn in Chiral Limit

The formal equivalence of the mass-gap equation (3.8) and the BSE (3.9) for a pseudoscalar meson in the chiral limit [12] will now be demonstrated by adapting them to a non-perturbative gluon exchange propagator [28] with an arbitrary confining form  $D(k)$  (not just the perturbative form  $k^{-2}$ ). The SDE, eq.(3.8), after replacing the color factor  $\lambda_1 \cdot \lambda_2/4$  by its Casimir value  $4/3$ , and a relabelling of symbols [28], now reads

$$\Sigma(p) = \frac{4}{3}i(2\pi)^{-4} \int d^4k D(\hat{k})\gamma\hat{\eta}S'_F(p-k)\gamma\hat{\eta}; \quad (3.10)$$

$S'_F$  is the full propagator related to the mass operator  $\Sigma(p)$  by

$$\Sigma(p) + i\gamma \cdot p = S_F^{-1}(p) = A(p^2)[i\gamma \cdot p + m(p^2)] \quad (3.11)$$

thus defining the mass function  $m(p^2)$  in the chiral limit  $m_c = 0$ . In the same way, eq.(3.9) for the vertex function  $\Gamma_H$  of a  $q\bar{q}$  hadron ( $H$ ) of 4-momentum  $P_\mu$  made up of quark 4-momenta  $p_{1,2} = P/2 \pm q$  reads as

$$\Gamma_H(q, P) = -\frac{4}{3}i(2\pi)^{-4} \int d^4q' D(\hat{q}-\hat{q}')\gamma\hat{\eta}S_F(q'+P/2)\Gamma_H(q', P)S_F(q'-P/2)\gamma\hat{\eta} \quad (3.12)$$

The complete equivalence of (3.10) and (3.12) for the pion case in the chiral limit  $P_\mu \rightarrow 0$  is easily established. Indeed, with the self-consistent ansatz  $\Gamma_H = \gamma_5 \Gamma(q)$ , eq.(3.12) simplifies to

$$\Gamma(q) = \frac{4}{3}i(2\pi)^{-4} \int d^4k \gamma \dot{\eta} S'_F(k-q) \Gamma(q-k) S'_F(q-k) \gamma \dot{\eta} D(\hat{k}) \quad (3.13)$$

where the replacement  $q' = q - k$  has been made. Substitution for  $S'_F$  from (3.11) in (3.13) gives

$$\Gamma(p) = -\frac{4}{3}i(2\pi)^{-4} \int d^4k \frac{D(\hat{k})\Gamma(p-k)}{A^2(p-k)(m^2((p-k)^2) + (p-k)^2)} \quad (3.14)$$

where we have relabelled  $q \rightarrow p$ . On the other hand substituting for  $S'_F$  (3.11) in (3.10) gives for the mass term of  $\Sigma(p)$  the result

$$A(p^2)m(p^2) = -\frac{4}{3}i(2\pi)^{-4} \int d^4k \frac{D(\hat{k})A(q')m(q'^2)}{A^2(q')(m^2(q'^2) + q'^2)} \quad (3.15)$$

where  $q' = p - k$ . A comparison of (3.14) and (3.15) shows their equivalence with the identification  $\Gamma(q) = A(q)m(q^2)$ , i.e. the identity of the vertex and mass functions in the chiral limit, provided  $A = 1$ , (which corresponds to the Landau gauge; see [32]). Although obtained here in the context of MYTP [26] this result is independent of this ansatz. A more explicit gauge theoretic derivation of the equations for the self-energy and vertex functions is given in [32].

### 3.3 Dynamical Mass As $DB\chi S$ Solution of SDE

We end this Section with the definition of the ‘dynamical’ mass function of the quark in the chiral limit ( $M_\pi = 0$ ) of the pion-quark vertex function  $\Gamma(\hat{q})$ , in the 3D-4D BSE framework [24,28]. The logic of this follows from the BSE-SDE formalism outlined above, eqs.(3.10-15), for the connection between eq.(3.15) for  $m(p)$  and eq.(3.14) for  $\Gamma(q)$  in the limit of zero mass of the pseudoscalar. So, setting  $M = 0$  in the (unnormalized)  $Hq\bar{q}$  vertex function  $\Gamma_H$  this quantity may be identified with the mass function  $m(\hat{p})$ , in the limit  $P_\mu = 0$ , where  $p_\mu$  is the 4-momentum of either quark; (note the appearance of the ‘hatted’ momentum). The result is [28,32]

$$m(\hat{p}) = [\omega(\hat{p}); \sqrt{2}p.n] \frac{m_q^2 + \hat{p}^2}{m_q^2} \phi(\hat{p}) \quad (3.16)$$

under CIA and Cov LF respectively. The normalization is such that in the low momentum limit, the constituent  $u\bar{d}$  mass  $m_q$  is recovered under CIA [28], while the corresponding ‘mass’ under Cov LF is  $p_+$  [32].

## 4 Null-Plane Preliminaries

The Weinberg infinite momentum method received a major boost through an understanding of Bjorken scaling [47a] in deep inelastic scattering, as well of the Feynman parton picture [47b] in the same process. The similarity of the  $P \rightarrow \text{inf}$  and the null-plane descriptions became clear with the demonstration by Susskind [48] of the  $U(2)$  structure of the LF/NP language wherein the role of ‘mass’ is played by the combination  $p_+ = p_0 + p_3$ , and subsequently a more complete formulation of null-plane dynamics by Kogut and Soper [49] within the Hamiltonian formalism in the context of field theory.

In a different direction, efforts were made to extend the Lorentz contraction ideas to finite momentum frames, designed to bring out the effect of Lorentz contraction on cluster form factors as a result of motion [50]. In the respect, the role of the Breit frame is particularly interesting since it gives the best possible overlap [50b] between the initial and final clusters. The Lorentz contraction factors [50] in turn are the key to an understanding of ‘dimensional scaling’ [51], especially in a ‘symmetrized’ version [50d] of the Breit frame [50b] which exactly reproduces the



correct ‘power counting’ [51]. And the Weinberg result [18] is duly recovered in the  $P \rightarrow \text{inf}$  limit, giving rise to the null-plane dynamics; (for more details of these results, see [38]). A more complete formulation of LF/NP dynamics, albeit with a *finite* number of degrees of freedom was given by Leutwyler-Stern [21] which comes closest to the original Dirac [20] spirit, and is summarized below.

#### 4.1 Leutwyler-Stern Formalism

Leutwyler-Stern [21] employed a Hamiltonian approach for investigating the properties of a relativistic 2-body system with a *finite* number of d.o.f.’s, and postulating a general criterion of ‘covariance’ in the form of an operator relation among the mass and spin operators of the system. Their formalism is based on the maximum ‘stability group’ (in the Dirac [20] sense) for the null-plane surface  $x_0 + x_3 = 0$ , which gives rise to the following seven ‘kinematical’ generators [38]:

$$\begin{aligned} \mathcal{K} &= (P_1, P_2, P_+, E_1, E_2, K_3, J_3); \quad 2P_{\pm} = P_0 \pm P_3; \\ 2E_1 &= K_1 + J_2, \quad 2E_2 = K_2 - J_1; \quad J_i = \frac{1}{2}\epsilon_{ijk}M_{jk}; \quad K_i = M_{0i}. \end{aligned} \quad (4.1)$$

The matrix elements of  $\mathcal{K}$  form a closed algebra ( $r, s = 1, 2$ ):

$$\begin{aligned} [K_3, E_r] &= -iE_r; \quad [K_3, P_+] = -iP_+; \quad [J_3, E_r] = +i\epsilon_{rs}E_s; \\ [J_3, P_r] &= +i\epsilon_{rs}P_s; \quad [E_r, P_s] = -i\delta_{rs}P_+ \end{aligned} \quad (4.2)$$

On the other hand, there are 3 ‘Hamiltonians’ which can be chosen in one of several different ways. To that end, it is necessary to introduce certain rotation operators  $I_i$  defined by  $I_i |n\rangle = J_i |n\rangle$ , on a rest system  $|n\rangle$ , but extended to states  $|\mathbf{p}, n\rangle$  defined by

$$|\mathbf{p}, n\rangle = \text{Exp}[-i\beta_1 E_1 - i\beta_2 E_2 - i\beta_3 K_3] |n\rangle; \quad \beta_r = p_r/p_+; \quad \beta_3 = \ln(2p_+/M) \quad (4.3)$$

such that  $\mathbf{I}$  commutes with the algebra of  $\mathcal{K}$ . More explicitly,

$$\begin{aligned} I_i |\mathbf{p}, n\rangle &= \text{Exp}[-i\beta_1 E_1 - i\beta_2 E_2 - i\beta_3 K_3] J_i |n\rangle; \\ [I_i, I_j] &= i\epsilon_{ijk} I_k; \quad [J_i, I_i] = 0. \end{aligned} \quad (4.4)$$

In particular,

$$I_3 = J_3 + (E_1 P_2 - E_2 P_1)/P_+ = (W_0 + W_3)/P_+; \quad W_\mu = \frac{1}{2}\epsilon_{\mu\nu\alpha\beta} P^\nu M^{\alpha\beta}. \quad (4.5)$$

$W_\mu$  is the Pauli-Lubanski operator, and  $MI_r = W_r - P_r W_+/P_+$ , where  $r = 1, 2$ . One thus arrives at the ‘dynamical group’  $D = (M, I_i)$ , or  $(M^2, MI_i)$ , which has the structure of  $U(2)$  [48], because of (3.4).  $D$  is really a 3-member group, since  $I_3$  already belongs to  $\mathcal{K}$  by virtue of (3.5). For a particular significance of  $\mathbf{I}$ , note its connection with the non-relativistic Galilei-invariant algebra generated by the momentum  $\mathbf{P}$  and Galilei boosts  $\mathbf{K}$ , viz.,

$$\mathbf{I} = \mathbf{J} + m^{-1}\mathbf{K} \times \mathbf{P}; \quad (m = \text{mass}) \quad (4.6)$$

Now Galilei invariance of a system is equivalent to the condition that the corresponding dynamical algebras constitute a unitary representation of  $U(2)$ . In the relativistic case, there is a superficial similarity to the  $U(2)$  structure of  $D$ , but unlike the NR case, only the component  $I_3$  is now ‘kinematical’, by virtue of (3.5), while  $I_{1,2}$  are ‘dynamical’, and do not have explicit representations by themselves. L-S [21] sought to bridge this gap by imposing the ‘covariance’ requirement in the form of an ‘angular condition’ for the operators  $I_i$  as follows:

$$x_1 M I_1 + x_2 M I_2 + x_L I_3; \quad x_L = P_1 x_1 + p_2 x_2 + P_+ x_- \quad (4.7)$$

which can be shown to be valid on the null-plane  $x_+ = x_0 + x_3 = 0$  [38].

The L-S formalism [21] provides a compact support to the longitudinal momentum  $z$  of 2-particle system with constituent 4-momenta  $p_{1,2}$ :

$$2zP_+ = p_{1+} - p_{2+}; \quad P_+ = p_{1+} + p_{2+}; \quad -\frac{1}{2} \leq x \leq +\frac{1}{2}. \quad (4.8)$$

The internal wave function  $\phi$  is defined by

$$\langle \mathbf{p}_1, \mathbf{p}_2 | \mathbf{P}, \phi \rangle = 2P_+ \delta^3(\mathbf{p}_1 + \mathbf{p}_2 - \mathbf{P}) \phi(\mathbf{q}_\perp, x) \quad (4.9)$$

where  $\phi$  is a matrix in helicity space  $(h', h'')$ , with the norm:

$$\langle \phi | \phi \rangle = \frac{1}{2} \int \frac{d^2 q_\perp dz}{\frac{1}{4} - x^2} \times \sum_{h' h''} |\phi_{h' h''}(\mathbf{q}_\perp, z)|^2 \quad (4.10)$$

The L-S [21] structure formally allows the introduction of a 3-vector  $\mathbf{q}$  and angular momentum  $\mathbf{L}$  for the internal motion of a composite of mass  $M$  with (equal) constituent masses  $m_q$  as

$$\mathbf{q} = (\mathbf{q}_\perp, Mx); \quad \mathbf{L} = -i\mathbf{q} \times \nabla_q \quad (4.11)$$

with the phase space

$$\frac{d^2 q_\perp dz}{\frac{1}{4} - x^2} = \frac{4d^3 \mathbf{q}}{M}; \quad M^2 = 4(m_q^2 + \mathbf{q}^2). \quad (4.12)$$

With these definitions of  $\mathbf{q}$  and  $\mathbf{L}$ , the theory formally preserves the concept of  $L$ -invariance of a  $q\bar{q}$  system despite the apparently asymmetric treatment meted out to the transverse and longitudinal components of 3-momenta in the NP or  $P = \inf$  formalism [18, 25, 48-9]. This invariance can be traced to angular condition (4.7). Incidentally, the historic FKR model [39], despite its other defects, was found by L-S [21] to satisfy the angular condition (4.7).

An alternative but more pedagogical recipe to achieve the same end was given in [38] via the simpler condition  $P.q = 0$ , to be consistently imposed between the total ( $P_\mu$ ) and relative ( $q_\mu$ ) 4-momenta of a  $q\bar{q}$  system, as outlined in subsection 4.2 below [38].

## 4.2 An Alternative "Angular Condition" $P.q = 0$

For unequal masses  $m_{1,2}$  of the (quark) constituents with 4-momenta  $p_{1,2}$ , the total ( $P$ ) and relative ( $q$ ) 4-momenta are given by the Wight-Gaerding definitions [52]

$$p_{1,2} = \hat{m}_{1,2} P \pm q; \quad 2\hat{m}_{1,2} = 1 \pm \frac{m_1^2 - m_2^2}{M^2} \quad (4.13)$$

where  $M = \sqrt{-P^2}$  is the composite mass. The condition  $P\dot{q} = 0$  is satisfied on the mass shells  $m_{1,2}^2 + p_{1,2}^2 = 0$  of the respective constituents, by virtue of the Wightman-Gaerding definition (4.13).

To link the condition  $P.q = 0$  with the construction of an effective 3-vector in the null-plane language, so as to preserve the invariance of the angular momentum concept, note that this condition translates to the relation  $q_- = -q_+ M^2 / P_+^2$  which expresses the component  $q_-$  in terms of  $q_+$ , in a frame  $P_\perp = 0$ , since in this frame,  $P_+ P_- = M^2$  on the mass shell of the composite. [The collinearity condition is not a restriction for a two-body system]. This relation then allows a definition of the 3-momentum  $\mathbf{q}$  with the components  $(\mathbf{q}_\perp, q_3)$ , with  $q_3 = M q_+ / P_+$ , which preserves the meaning of  $\mathbf{L}$  in the sense of L-S [21], together with NP covariance. For any other internal 4-vector  $A_\mu$  for the composite, a similar 3-vector  $\mathbf{A}$  may be defined as  $(\mathbf{A}_\perp, A_3)$ , with  $A_3 = M A_+ / P_+$ , via the condition  $A\dot{P} = 0$ . Examples of  $A_\mu$  are polarization vectors, Dirac matrices, etc. Using these techniques, null-plane wave functions of the L-S type [21] have been constructed and applied to hadronic processes via quark loops [34]. A more formal mathematical basis for this prescription comes from MYTP [26] on the covariant null-plane [37]; see Sect.5 below.

## 5 3D-4D BSE Under MYTP: Scalar/Fermion Quarks

We now come to our objectives (B) and (C), viz., 3D-4D interlinkage of BS amplitudes brought about by MYTP [26], and a unified treatment of  $q\bar{q}$  [24,37] and  $qqq$  [53] systems under MYTP conditions. The full calculational details with 4-fermion couplings via gluonic propagators have been collected in a recent review [32]. However, to bring out the basic mathematical structures, we shall derive the 3D-4D interconnection with spinless quarks for 2- and 3-quark systems in this and the next sections respectively. Further, we shall consider two MYTP methods in parallel for comparison: i) Covariant Instantaneity Ansatz (CIA) [24-25]; ii) Covariant LF/NP Ansatz (Cov LF) [37], to bring out the structural identity of the resulting BSE's for a  $q\bar{q}$  system. This will be followed by a reconstruction of the 4D BS vertex functions for both types [24, 37] as basic building blocks for 4D quark loop calculations.

### 5.1 3D-4D BSE Under CIA: Spinless Quarks

For a self-contained presentation, with unequal mass kinematics [24], let the quark constituents of masses  $m_{1,2}$  and 4-momenta  $p_{1,2}$  interact to produce a composite hadron of mass  $M$  and 4-momentum  $P_\mu$ . The relative 4-momentum  $q_\mu$  is related to these by

$$p_{1,2} = \hat{m}_{1,2}P \pm q; \quad P^2 = -M^2; \quad 2\hat{m}_{1,2} = 1 \pm (m_1^2 - m_2^2)/M^2 \quad (5.1)$$

These Wightman-Garding definitions [52] of the fractional momenta  $\hat{m}_{1,2}$  ensure that  $q.P = 0$  on the mass shells  $m_i^2 + p_i^2 = 0$  of the constituents, though not off-shell. Now define  $\hat{q}_\mu = q_\mu - q.P P_\mu / P^2$  as the relative momentum *transverse* to the hadron 4-momentum  $P_\mu$  which automatically gives  $\hat{q}.P \equiv 0$ , for all values of  $\hat{q}_\mu$ . If the BSE kernel  $K$  for the 2 quarks is a function of only these transverse relative momenta, viz.  $K = K(\hat{q}, \hat{q}')$ , this is called the ‘‘Cov. Inst. Ansatz (CIA)’’ [24] which accords with MYTP [26]. For two scalar quarks with inverse propagators  $\Delta_{1,2}$ , this ansatz gives rise to the following BSE for the wave fn  $\Phi(q, P)$  [24]:

$$i(2\pi)^4 \Delta_1 \Delta_2 \Phi(q, P) = \int d^4 q' K(\hat{q}, \hat{q}') \Phi(q', P); \quad \Delta_{1,2} = m_{1,2}^2 + p_{1,2}^2 \quad (5.2)$$

The quantities  $m_{1,2}$  are the ‘constituent’ masses which are strictly momentum dependent since they contain the mass function  $m(p)$  [12,28], but may be regarded as constant for low energy phenomena:  $m(p) \cong m(0)$ . Further, under CIA,  $m(p) = m(\hat{p})$ , a momentum-dependence which is governed by the  $DB\chi S$  condition [28] (see below).

To make a 3D reduction of eq.(5.2), define the 3D wave function  $\phi(\hat{q})$  in terms of the longitudinal momentum  $M\sigma$  as

$$\phi(\hat{q}) = \int M d\sigma \Phi(q, P); \quad M\sigma = Mq.P / P^2 \quad (5.3)$$

using which, eq.(5.2) may be recast as

$$i(2\pi)^4 \Delta_1 \Delta_2 \Phi(q, P) = \int d^3 \hat{q}' K(\hat{q}, \hat{q}') \phi(\hat{q}'); \quad d^4 q' \equiv d^3 \hat{q}' M d\sigma' \quad (5.4)$$

Next, divide out by  $\Delta_1 \Delta_2$  in (5.4) and use once again (5.3) to reduce the 4D BSE form (5.4) to the 3D form

$$(2\pi)^3 D(\hat{q}) \phi(\hat{q}) = \int d^3 \hat{q}' K(\hat{q}, \hat{q}') \phi(\hat{q}'); \quad \frac{2i\pi}{D(\hat{q})} \equiv \int \frac{M d\sigma}{\Delta_1 \Delta_2} \quad (5.5)$$

Here  $D(\hat{q})$  is the 3D denominator function associated with the like wave function  $\phi(\hat{q})$ . The integration over  $d\sigma$  is carried out by noting pole positions of  $\Delta_{1,2}$  in the  $\sigma$ -plane, where

$$\Delta_{1,2} = \omega_{1,2}^2 - M^2(\hat{m}_{1,2} \pm \sigma)^2; \quad \omega_{1,2}^2 = m_{1,2}^2 + \hat{q}^2 \quad (5.6)$$

The pole positions are given for  $\Delta_{1,2} = 0$  respectively by

$$M(\sigma + \hat{m}_1) = \pm \omega_1 \mp i\epsilon; \quad M(\sigma - \hat{m}_2) = \pm \omega_2 \mp i\epsilon \quad (5.7)$$

where the  $(\pm)$  indices refer to the lower/upper halves of the  $\sigma$ - plane. The final result for  $D(\hat{q})$  is expressible symmetrically [24]:

$$D(\hat{q}) = M_\omega D_0(\hat{q}); \quad \frac{2}{M_\omega} = \frac{\hat{m}_1}{\omega_1} + \frac{\hat{m}_2}{\omega_2} \quad (5.8)$$

$$\frac{1}{2}D_0(\hat{q}) = \hat{q}^2 - \frac{\lambda(m_1^2, m_2^2, M^2)}{4M^2}; \quad \lambda = M^4 - 2M^2(m_1^2 + m_2^2) + (m_1^2 - m_2^2)^2 \quad (5.9)$$

The crucial thing for MYTP[26] is now to observe the *equality* of the RHS of eqs (5.4) and (5.5), thus leading to an *exact interconnection* between the 3D and 4D BS wave functions [24]:

$$\Gamma(\hat{q}) \equiv \Delta_1 \Delta_2 \Phi(q, P) = \frac{D(\hat{q}\phi(\hat{q}))}{2i\pi} \quad (5.10)$$

Eq.(5.10) determines the hadron-quark vertex function  $\Gamma(\hat{q})$  as a product  $D\phi$  of the 3D denominator and wave functions, satisfying a relativistic 3D Schroedinger-like equation (5.5). The simultaneous appearance of the 3D form (5.5) and the 4D form (5.4), leading to their interconnection (5.10), reveals a two-tier character: The 3D form (5.5) gives the basis for making contact with the 3D spectra [13], while the reconstructed 4D wave (vertex) function (5.10) in terms of 3D ingredients  $D$  and  $\phi$  can be used for 4D quark-loop integrals in the standard Feynman fashion. Note that the vertex function  $\Gamma = D\phi/(2i\pi)$  has a general structure, independent of the details of the input kernel  $K$ . Further, the  $D$ -function, eq.(5.8), is universal and well-defined off the mass shell of either quark. The 3D wave function  $\phi$  is admittedly model- dependent, but together with  $D(\hat{q})$ , it controls the 3D spectra via (5.5), thus offering a direct experimental check on its structure. Both functions depend on the single 3D Lorentz-covariant quantity  $\hat{q}^2$  whose most important property is its positive definiteness for time-like hadron momenta ( $M^2 > 0$ ).

## 5.2 Cov LF/NP for 3D-4D BSE: Fermion Quarks

As a preliminary to defining a 3D support to the BS kernel on the light-front (LF/NP), on the lines of CIA [24], a covariant LF/NP orientation [37] may be represented by the 4-vector  $n_\mu$ , as well as its dual  $\tilde{n}_\mu$ , obeying the normalizations  $n^2 = \tilde{n}^2 = 0$  and  $n \cdot \tilde{n} = 1$ . In the standard NP scheme (in euclidean notation), these quantities are  $n = (001; -i)/\sqrt{2}$  and  $\tilde{n} = (001; i)/\sqrt{2}$ , while the two other perpendicular directions are collectively denoted by the subscript  $\perp$  on the concerned momenta. We shall try to maintain the  $n$ -dependence of various momenta to ensure explicit covariance; and to keep track of the usual NP notation  $p_\pm = p_0 \pm p_3$ , our covariant notation is normalized to the latter as  $p_+ = n \cdot p\sqrt{2}$ ;  $p_- = -\tilde{n} \cdot p\sqrt{2}$ , while the perpendicular components continue to be denoted by  $p_\perp$  in both notations.

In the same notation as for CIA [24], the 4th component of the relative momentum  $q = \hat{m}_2 p_1 - \hat{m}_1 p_2$ , that should be eliminated for obtaining a 3D equation, is now proportional to  $q_n \equiv \tilde{n} \cdot q$ , as the NP analogue [37] of  $P \cdot q P / P^2$  in CIA [24], where  $P = p_1 + p_2$  is the total 4-momentum of the hadron. However the quantity  $q - q_n n$  is still only  $q_\perp$ , since its square is  $q^2 - 2n \cdot q \tilde{n} \cdot q$ , as befits  $q_\perp^2$  (readily checked against the 'special' NP frame). We still need a third component  $p_3$ , for which the correct definition turns out to be [37]  $q_{3\mu} = z P_n n_\mu$ , where  $P_n = P \cdot \tilde{n}$  and  $z = q \cdot n / P \cdot n$ , which checks with  $\hat{q}^2 = q_\perp^2 + z^2 M^2$ . We now collect the following definitions/results:

$$\begin{aligned} q_\perp &= q - q_n n; \quad \hat{q} = q_\perp + x P_n n; \quad x = q \cdot n / P \cdot n; \quad P^2 = -M^2; \\ q_n, P_n &= \tilde{n} \cdot (q, P); \quad \hat{q} \cdot n = q \cdot n; \quad \hat{q} \cdot \tilde{n} = 0; \quad P_\perp \cdot q_\perp = 0; \\ P \cdot q &= P_n q \cdot n + P \cdot n q_n; \quad P \cdot \hat{q} = P_n q \cdot n; \quad \hat{q}^2 = q_\perp^2 + M^2 x^2 \end{aligned} \quad (5.11)$$

Now in analogy to CIA, the reduced 3D BSE (wave-fn  $\phi$ ) may be derived from the 4D BSE (5.2) for spinless quarks (wave-fn  $\Phi$ ) when its kernel  $K$  is *decreed* to be independent of the component  $q_n$ , i.e.,  $K = K(\hat{q}, \hat{q}')$ , with  $\hat{q} = (q_\perp, P_n n)$ , in accordance with MYTP [26] condition imposed on the null-plane (NP), so that  $d^4 q = d^2 q_\perp dq_3 dq_n$ . Now define a 3D wave-fn  $\phi(\hat{q}) = \int dq_n \Phi(q)$ , as the

CNPA counterpart of the CIA definition (5.3), and use this result on the RHS of (5.2) to give

$$i(2\pi)^4 \Phi(q) = \Delta_1^{-1} \Delta_2^{-1} \int d^3 \hat{q}' K(\hat{q}, \hat{q}') \phi(\hat{q}') \quad (5.12)$$

which is formally the same as eq.(5.4) for CIA above. Now integrate both sides of eq.(5.12) w.r.t.  $dq_n$  to give a 3D BSE in the variable  $\hat{q}$ :

$$(2\pi)^3 D_n(\hat{q}) \phi(\hat{q}) = \int d^2 q_{\perp}' dq_3' K(\hat{q}, \hat{q}') \phi(\hat{q}') \quad (5.13)$$

which again corresponds to the CIA eq.(5.5), except that the function  $D_n(\hat{q})$  is now defined by

$$\int dq_n \Delta_1^{-1} \Delta_2^{-1} = 2\pi i D_n^{-1}(\hat{q}) \quad (5.14)$$

and may be obtained by standard NP techniques [38] (Chaps 5-7) as follows. In the  $q_n$  plane, the poles of  $\Delta_{1,2}$  lie on opposite sides of the real axis, so that only *one* pole will contribute at a time. Taking the  $\Delta_2$ -pole, which gives

$$2q_n = -\sqrt{2}q_{-} = \frac{m_2^2 + (q_{\perp} - \hat{m}_2 P)^2}{\hat{m}_2 P \cdot n - q \cdot n} \quad (5.15)$$

the residue of  $\Delta_1$  works out, after a routine simplification, to just  $2P \cdot q = 2P \cdot n q_n + 2P_n q \cdot n$ , after using the collinearity condition  $P_{\perp} \cdot q_{\perp} = 0$  from (5.11). And when the value (5.15) of  $q_n$  is substituted in (5.14), one obtains (with  $P_n P \cdot n = -M^2/2$ ):

$$D_n(\hat{q}) = 2P \cdot n (\hat{q}^2 - \frac{\lambda(M^2, m_1^2, m_2^2)}{4M^2}); \quad \hat{q}^2 = q_{\perp}^2 + M^2 x^2; \quad x = q \cdot n / P \cdot n \quad (5.16)$$

Now a comparison of (5.12) with (5.13) relates the 4D and 3D wave-fns:

$$2\pi i \Phi(q) = D_n(\hat{q}) \Delta_1^{-1} \Delta_2^{-1} \phi(\hat{q}) \quad (5.17)$$

as the Cov LF counterpart of (5.10) which is valid near the bound state pole. The BS vertex function now becomes  $\Gamma = D_n \times \phi / (2\pi i)$ . This result, though dependent on the LF/NP orientation, is nevertheless formally *covariant*, and closely corresponds to the pedagogical result of the old LF/NP formulation [38], with  $D_n \Leftrightarrow D_+$ .

A 3D equation similar to the covariant eq.(5.13) above, also obtains in alternative LF formulations such as in Kadychevsky-Karmanov [22b] (see their eq.(3.48)). However the *independent* 4-vector  $\tilde{n}_{\mu}$  (which has no counterpart in [22b]), makes this a manifestly covariant 4D formulation without need for explicit Lorentz transformations [22b]. The 'angular condition' [21] is also trivially satisfied by the effective 3-vector  $\hat{q}_{\mu}$  appearing in the 3D BSE (5.13). A more important contrast from other null-plane approaches is that the inverse process of *reconstruction* of the 4D hadron-quark vertex, eq.(5.17), has no counterpart in them [22-23], as these are basically 3D oriented. *not vice versa*.

For fermion quarks with gluonic propagators, the MYTP formulation needs no new principles, except for certain technical details involving slight modifications [54] of the BSE structure for easier handling; see [32] for detailed steps. The full 4D wave function  $\Psi(P, q)$  may be expressed as a 4x4 matrix [38,32]:

$$\Psi(P, q) = S_F(p_1) \Gamma(\hat{q}) \gamma_D S_F(-p_2); \quad \Gamma(\hat{q}) = N_H [1; P_n/M] D(\hat{q}) \phi(\hat{q}) / 2i\pi \quad (5.18)$$

where  $\gamma_D$  is a Dirac matrix which equals  $\gamma_5$  for a P-meson,  $i\gamma_{\mu}$  for a V-meson,  $i\gamma_{\mu}\gamma_5$  for an A-meson, etc. The factors in square brackets stand for CIA and Cov LF values respectively.  $N_H$  represents the hadron normalization.

## 6 The $qqq$ BSE: 3D-4D Interlinkage

We now come to the aspect of MYTP [26] that governs the inter-relation of 3D and 4D Bethe-Salpeter amplitudes for 3-body ( $qqq$ )-systems, in keeping with a perceived ‘duality’ between meson ( $q\bar{q}$ ) and baryon ( $qqq$ ) systems which necessitates a parallel treatment between them. In this respect a fairly comprehensive review of baryon dynamics as a 3-body relativistic system with full permutation symmetries in all relevant degrees of freedom [55] has been given recently [32]. These include: A detailed correspondence [56] between  $qqq$  and quark-diquark wave functions; Complex HO techniques for the  $qqq$  problem [57]; fermionic  $qqq$  BSE with the same gluon propagator for pair  $qq$  interactions [29] as employed for  $q\bar{q}$  systems [28], except for reduction by half due to the color factor; and Green’s function methods for 3D reduction of the 4D BSE form, plus reconstruction of the 4D  $qqq$  wave function [53], on the lines of the  $q\bar{q}$  problem [24]. Within the formalistic scope of this Article however, we shall merely dwell on the last item, viz., Green’s fn techniques [53] for a 3D reduction of the 4D BSE, plus *reconstruction* of the 4D wave function, for a  $qqq$  system for three identical spinless quarks, keeping in forefront the issue of *connectedness* [58] in a 3-particle amplitude whose signal is the *absence of any  $\delta$ -function* in its structure; (for a detailed perspective, see [32]).

### 6.1 Two-Quark Green’s Function Under CIA

As a warm up to the method of Green’s functions (G-fns), we first derive the 3D-4D interconnection for the corresponding G-fns for 2-particle scattering of two identical spinless particles, before moving on to the 3-body problem in the next 2 subsections. For simplicity we shall consider the G-fns near the bound state poles, so that the inhomogeneous terms may be dropped. In the notation and phase convention of Section 5, the 4D  $qq$  Green’s fn  $G(p_1 p_2; p_1' p_2')$  near a *bound* state satisfies a 4D BSE (no inhomogeneous term):

$$i(2\pi)^4 G(p_1 p_2; p_1' p_2') = \Delta_1^{-1} \Delta_2^{-1} \int dp_1'' dp_2'' K(p_1 p_2; p_1'' p_2'') G(p_1'' p_2''; p_1' p_2'); \quad (6.1.1)$$

where

$$\Delta_1 = p_1^2 + m_q^2, \quad (6.1.2)$$

and  $m_q$  is the mass of each quark. Now using the relative 4-momentum  $q = (p_1 - p_2)/2$  and total 4-momentum  $P = p_1 + p_2$  (similarly for the other sets), and removing a  $\delta$ -function for overall 4-momentum conservation, from each of the  $G$ - and  $K$ - functions, eq.(6.1.1) reduces to the simpler form

$$i(2\pi)^4 G(q, q') = \Delta_1^{-1} \Delta_2^{-1} \int d\hat{q}'' M d\sigma'' K(\hat{q}, \hat{q}'') G(q'', q') \quad (6.1.3)$$

where  $\hat{q}_\mu = q_\mu - \sigma P_\mu$ , with  $\sigma = (q \cdot P)/P^2$ , is effectively 3D in content (being orthogonal to  $P_\mu$ ). Here we have incorporated the ansatz of a 3D support for the kernel  $K$  (independent of  $\sigma$  and  $\sigma'$ ), and broken up the 4D measure  $dq''$  arising from (6.1.1) into the product  $d\hat{q}'' M d\sigma''$  of a 3D and a 1D measure respectively. We have also suppressed the 4-momentum  $P_\mu$  label, with  $(P^2 = -M^2)$ , in the notation for  $G(q, q')$ .

Now define the fully 3D Green’s function  $\hat{G}(\hat{q}, \hat{q}')$  as [53]

$$\hat{G}(\hat{q}, \hat{q}') = \int \int M^2 d\sigma d\sigma' G(q, q') \quad (6.1.4)$$

and two (hybrid) 3D-4D Green’s functions  $\tilde{G}(\hat{q}, q')$ ,  $\tilde{G}(q, \hat{q}')$  as

$$\tilde{G}(\hat{q}, q') = \int M d\sigma G(q, q'); \quad \tilde{G}(q, \hat{q}') = \int M d\sigma' G(q, q'); \quad (6.1.5)$$

Next, use (6.1.5) in (6.1.3) to give

$$i(2\pi)^4 \tilde{G}(q, \hat{q}') = \Delta_1^{-1} \Delta_2^{-1} \int d\hat{q}'' K(\hat{q}, \hat{q}'') \tilde{G}(q'', \hat{q}') \quad (6.1.6)$$

Now integrate both sides of (6.1.3) w.r.t.  $Md\sigma$  and use the result

$$\int M d\sigma \Delta_1^{-1} \Delta_2^{-1} = 2\pi i D^{-1}(\hat{q}); \quad D(\hat{q}) = 4\hat{\omega}(\hat{\omega}^2 - M^2/4); \quad \hat{\omega}^2 = m_q^2 + \hat{q}^2 \quad (6.1.7)$$

to give a 3D BSE w.r.t. the variable  $\hat{q}$ , while keeping the other variable  $q'$  in a 4D form:

$$(2\pi)^3 \tilde{G}(\hat{q}, q') = D^{-1} \int d\hat{q}'' K(\hat{q}, \hat{q}'') \tilde{G}(\hat{q}'', q') \quad (6.1.8)$$

A comparison of (6.1.3) with (6.1.8) gives the desired connection between the full 4D  $G$ -function and the hybrid  $\tilde{G}(\hat{q}, q')$ -function:

$$2\pi i G(q, q') = D(\hat{q}) \Delta_1^{-1} \Delta_2^{-1} \tilde{G}(\hat{q}, q') \quad (6.1.9)$$

Again, the symmetry of the left hand side of (6.1.9) w.r.t.  $q$  and  $q'$  allows rewriting the right hand side with the roles of  $q$  and  $q'$  interchanged. This gives the dual form

$$2\pi i G(q, q') = D(\hat{q}') \Delta_1'^{-1} \Delta_2'^{-1} \tilde{G}(\hat{q}, \hat{q}') \quad (6.1.10)$$

which on integrating both sides w.r.t.  $Md\sigma$  gives

$$2\pi i \tilde{G}(\hat{q}, q') = D(\hat{q}') \Delta_1'^{-1} \Delta_2'^{-1} \hat{G}(\hat{q}, \hat{q}'). \quad (6.1.11)$$

Substitution of (6.1.11) in (6.1.9) then gives the symmetrical form

$$(2\pi i)^2 G(q, q') = D(\hat{q}) \Delta_1^{-1} \Delta_2^{-1} \tilde{G}(\hat{q}, \hat{q}') D(\hat{q}') \Delta_1'^{-1} \Delta_2'^{-1} \quad (6.1.12)$$

Finally, integrating both sides of (6.1.8) w.r.t.  $Md\sigma'$ , we obtain a fully reduced 3D BSE for the 3D Green's function:

$$(2\pi)^3 \hat{G}(\hat{q}, \hat{q}') = D^{-1}(\hat{q}) \int d\hat{q}'' K(\hat{q}, \hat{q}'') \hat{G}(\hat{q}'', \hat{q}') \quad (6.1.13)$$

Eq.(6.1.12) which is valid near the bound state pole, expresses the desired connection between the 3D and 4D forms of the Green's functions; and eq(6.1.13) is the determining equation for the 3D form. A spectral analysis can now be made for either of the 3D or 4D Green's functions in the standard manner, viz.,

$$G(q, q') = \sum_n \Phi_n(q; P) \Phi_n^*(q'; P) / (P^2 + M^2) \quad (6.1.14)$$

where  $\Phi$  is the 4D BS wave function. A similar expansion holds for the 3D  $G$ -function  $\hat{G}$  in terms of  $\phi_n(\hat{q})$ . Substituting these expansions in (6.1.12), one immediately sees the connection between the 3D and 4D wave functions in the form:

$$2\pi i \Phi(q, P) = \Delta_1^{-1} \Delta_2^{-1} D(\hat{q}) \phi(\hat{q}) \quad (6.1.15)$$

whence the BS vertex function becomes  $\Gamma = D \times \phi / (2\pi i)$  as found in [24]. We shall make free use of these results, taken as  $qq$  subsystems, for our study of the  $qqq$   $G$ -functions in subsects.6.2-3.

## 6.2 3D BSE Reduction for $qqq$ G-fn

As in the two-body case, and in an obvious notation for various 4-momenta (without the Greek suffixes), we consider the most general Green's function  $G(p_1 p_2 p_3; p_1' p_2' p_3')$  for 3-quark scattering *near the bound state pole* (for simplicity) which allows us to drop the various inhomogeneous terms from the beginning. Again we take out an overall delta function  $\delta(p_1 + p_2 + p_3 - P)$  from the  $G$ -function and work with *two* internal 4-momenta for each of the initial and final states defined as follows [54b]:

$$\sqrt{3}\xi_3 = p_1 - p_2; \quad 3\eta_3 = -2p_3 + p_1 + p_2 \quad (6.2.1)$$

$$P = p_1 + p_2 + p_3 = p_1' + p_2' + p_3' \quad (6.2.2)$$

and two other sets  $\xi_1, \eta_1$  and  $\xi_2, \eta_2$  defined by cyclic permutations from (6.2.1). Further, as we shall consider pairwise kernels with 3D support, we define the effectively 3D momenta  $\hat{p}_i$ , as well as the three (cyclic) sets of internal momenta  $\hat{\xi}_i, \hat{\eta}_i$ , ( $i = 1, 2, 3$ ) by [54b]:

$$\hat{p}_i = p_i - \nu_i P; \quad \hat{\xi}_i = \xi_i - s_i P; \quad \hat{\eta}_i = \eta_i - t_i P \quad (6.2.3)$$

$$\nu_i = (P \cdot p_i)/P^2; \quad s_i = (P \cdot \xi_i)/P^2; \quad t_i = (P \cdot \eta_i)/P^2 \quad (6.2.4)$$

$$\sqrt{3}s_3 = \nu_1 - \nu_2; \quad 3t_3 = -2\nu_3 + \nu_1 + \nu_2 \quad (+\text{cyclic permutations}) \quad (6.2.5)$$

The space-like momenta  $\hat{p}_i$  and the time-like ones  $\nu_i$  satisfy [54b]

$$\hat{p}_1 + \hat{p}_2 + \hat{p}_3 = 0; \quad \nu_1 + \nu_2 + \nu_3 = 1 \quad (6.2.6)$$

Strictly speaking, in the spirit of covariant instantaneity, we should have taken the relative 3D momenta  $\hat{\xi}, \hat{\eta}$  to be in the instantaneous frames of the concerned pairs, i.e., w.r.t. the rest frames of  $P_{ij} = p_i + p_j$ ; however the difference between the rest frames of  $P$  and  $P_{ij}$  is small and calculable [54b], while the use of a common 3-body rest frame ( $P = 0$ ) lends considerable simplicity and elegance to the formalism.

We may now use the foregoing considerations to write down the BSE for the 6-point Green's function in terms of relative momenta, on closely parallel lines to the 2-body case. To that end note that the 2-body relative momenta are  $q_{ij} = (p_i - p_j)/2 = \sqrt{3}\xi_k/2$ , where (ijk) are cyclic permutations of (123). Then for the reduced  $qqq$  Green's function, when the *last* interaction was in the (ij) pair, we may use the notation  $G(\xi_k \eta_k; \xi_k' \eta_k')$ , together with 'hat' notations on these 4-momenta when the corresponding time-like components are integrated out. Further, since the pair  $\xi_k, \eta_k$  is *permutation invariant* as a whole, we may choose to drop the index notation from the complete  $G$ -function to emphasize this symmetry as and when needed. The  $G$ -function for the  $qqq$  system satisfies, in the neighbourhood of the bound state pole, the following (homogeneous) 4D BSE for pairwise  $qq$  kernels with 3D support:

$$i(2\pi)^4 G(\xi\eta; \xi'\eta') = \sum_{123} \Delta_1^{-1} \Delta_2^{-1} \int d\hat{q}_{12}'' M d\sigma_{12}'' K(\hat{q}_{12}, \hat{q}_{12}'') G(\xi_3'' \eta_3''; \xi_3' \eta_3') \quad (6.2.7)$$

where we have employed a mixed notation ( $q_{12}$  versus  $\xi_3$ ) to stress the two-body nature of the interaction with one spectator at a time, in a normalization directly comparable with eq.(6.1.3) for the corresponding two-body problem. Note also the connections

$$\sigma_{12} = \sqrt{3}s_3/2; \quad \hat{q}_{12} = \sqrt{3}\hat{\xi}_3/2; \quad \hat{\eta}_3 = -\hat{p}_3, \quad \text{etc} \quad (6.2.8)$$

The next task is to reduce the 4D BSE (6.2.7) to a fully 3D form through a sequence of integrations w.r.t. the time-like momenta  $s_i, t_i$  applied to the different terms on the right hand side, *provided both* variables are simultaneously permuted. We now define the following fully 3D as well as mixed (hybrid) 3D-4D  $G$ -functions according as one or more of the time-like  $\xi, \eta$  variables are integrated out:

$$\hat{G}(\hat{\xi}\hat{\eta}; \hat{\xi}'\hat{\eta}') = \int \int \int \int ds dt ds' dt' G(\xi\eta; \xi'\eta') \quad (6.2.9)$$

which is  $S_3$ -symmetric.

$$\tilde{G}_{3\eta}(\xi\hat{\eta}; \xi'\hat{\eta}') = \int \int dt_3 dt_3' G(\xi\eta; \xi'\eta'); \quad (6.2.10)$$

$$\tilde{G}_{3\xi}(\hat{\xi}\eta; \hat{\xi}'\eta') = \int \int ds_3 ds_3' G(\xi\eta; \xi'\eta'); \quad (6.2.11)$$

The last two equations are however *not* symmetric w.r.t. the permutation group  $S_3$ , since both the variables  $\xi, \eta$  are not simultaneously transformed; this fact has been indicated in eqs.(8.2.10-11) by the suffix "3" on the corresponding (hybrid)  $\tilde{G}$ -functions, to emphasize that the 'asymmetry'



is w.r.t. the index "3". We shall term such quantities " $S_3$ -indexed", to distinguish them from  $S_3$ -symmetric quantities as in eq.(6.2.9). The full 3D BSE for the  $\hat{G}$ -function is obtained by integrating out both sides of (6.2.7) w.r.t. the  $st$ -pair variables  $ds; ds_j' dt; dt_j'$  (giving rise to an  $S_3$ -symmetric quantity), and using (6.2.9) together with (6.2.8) as follows:

$$(2\pi)^3 \hat{G}(\hat{\xi}\hat{\eta}; \hat{\xi}'\hat{\eta}') = \sum_{123} D^{-1}(\hat{q}_{12}) \int d\hat{q}_{12}'' K(\hat{q}_{12}, \hat{q}_{12}'') \hat{G}(\hat{\xi}''\hat{\eta}''; \hat{\xi}'\hat{\eta}') \quad (6.2.12)$$

This integral equation for  $\hat{G}$  which is the 3-body counterpart of (6.1.13) for a  $qq$  system in the neighbourhood of the bound state pole, is the desired 3D BSE for the  $qqq$  system in a *fully connected* form, i.e., free from delta functions. Now using a spectral decomposition for  $\hat{G}$

$$\hat{G}(\hat{\xi}\hat{\eta}; \hat{\xi}'\hat{\eta}') = \sum_n \phi_n(\hat{\xi}\hat{\eta}; P) \phi_n^*(\hat{\xi}'\hat{\eta}'; P) / (P^2 + M^2) \quad (6.2.13)$$

on both sides of (6.2.12) and equating the residues near a given pole  $P^2 = -M^2$ , gives the desired equation for the 3D wave function  $\phi$  for the bound state in the connected form:

$$(2\pi)^3 \phi(\hat{\xi}\hat{\eta}; P) = \sum_{123} D^{-1}(\hat{q}_{12}) \int d\hat{q}_{12}'' K(\hat{q}_{12}, \hat{q}_{12}'') \phi(\hat{\xi}''\hat{\eta}''; P) \quad (6.2.14)$$

Now the  $S_3$ -symmetry of  $\phi$  in the  $(\hat{\xi}_i, \hat{\eta}_i)$  pair is a very useful result for both the solution of (6.2.14) *and* for the reconstruction of the 4D BS wave function in terms of the 3D wave function (6.2.14), as is done in the subsect.6.3 below.

### 6.3 Reconstruction of 4D $qqq$ Wave Function

We now attempt to *re-express* the 4D  $G$ -function given by (6.2.7) in terms of the 3D  $\hat{G}$ -function given by (6.2.12), as the  $qqq$  counterpart of the  $qq$  results (6.1.12-13). To that end we adapt the result (6.1.12) to the hybrid Green's function of the (12) subsystem given by  $\tilde{G}_{3\eta}$ , eq.(6.2.10), in which the 3-momenta  $\hat{\eta}_3, \hat{\eta}_3'$  play a parametric role reflecting the spectator status of quark #3, while the *active* roles are played by  $q_{12}, q_{12}' = \sqrt{3}(\xi_3, \xi_3')/2$ , for which the analysis of subsect.6.1 applies directly. This gives

$$(2\pi i)^2 \tilde{G}_{3\eta}(\xi_3 \hat{\eta}_3; \xi_3' \hat{\eta}_3') = D(\hat{q}_{12}) \Delta_1^{-1} \Delta_2^{-1} \hat{G}(\hat{\xi}_3 \hat{\eta}_3; \hat{\xi}_3' \hat{\eta}_3') D(\hat{q}_{12}') \Delta_1'^{-1} \Delta_2'^{-1} \quad (6.3.1)$$

where on the right hand side, the 'hatted'  $G$ -function has full  $S_3$ -symmetry, although (for purposes of book-keeping) we have not shown this fact explicitly by deleting the suffix '3' from its arguments. A second relation of this kind may be obtained from (6.2.7) by noting that the 3 terms on its right hand side may be expressed in terms of the hybrid  $\tilde{G}_{3\xi}$  functions vide their definitions (6.2.11), together with the 2-body interconnection between  $(\xi_3, \xi_3')$  and  $(\hat{\xi}_3, \hat{\xi}_3')$  expressed once again via (6.3.1), but without the 'hats' on  $\eta_3$  and  $\eta_3'$ . This gives

$$\begin{aligned} (\sqrt{3}\pi i)^2 G(\xi_3 \eta_3; \xi_3' \eta_3') &= (\sqrt{3}\pi i)^2 G(\xi\eta; \xi'\eta') \\ &= \sum_{123} \Delta_1^{-1} \Delta_2^{-1} (\pi i \sqrt{3}) \int d\hat{q}_{12}'' M d\sigma_{12}'' K(\hat{q}_{12}, \hat{q}_{12}'') G(\xi_3'' \eta_3''; \xi_3' \eta_3') \\ &= \sum_{123} D(\hat{q}_{12}) \Delta_1^{-1} \Delta_2^{-1} \tilde{G}_{3\xi}(\hat{\xi}_3 \eta_3; \hat{\xi}_3' \eta_3') \Delta_1'^{-1} \Delta_2'^{-1} \end{aligned} \quad (6.3.2)$$

where the second form exploits the symmetry between  $\xi, \eta$  and  $\xi', \eta'$ .

At this stage, unlike the 2-body case, the reconstruction of the 4D Green's function is *not yet complete* for the 3-body case, as eq.(6.3.2) clearly shows. This is due to the *truncation* of Hilbert space implied in the ansatz of 3D support to the pairwise BSE kernel  $K$  which, while facilitating a 4D to 3D BSE reduction without extra charge, does *not* have the *complete* information to permit the *reverse* transition (3D to 4D) without additional assumptions. To fill up this gap in this

“inverse” mathematical problem, we look for a suitable ansatz for  $\tilde{G}_{3\xi}$  on the RHS of (6.3.2) in terms of *known* quantities, so that the reconstructed 4D  $G$ -function satisfies the 3D equation (6.2.12) exactly, as a check-point. We therefore seek a structure of the form

$$\tilde{G}_{3\xi}(\hat{\xi}_3\eta_3; \hat{\xi}_3'\eta_3') = \hat{G}(\hat{\xi}_3\hat{\eta}_3; \hat{\xi}_3'\hat{\eta}_3') \times F(p_3, p_3') \quad (6.3.3)$$

where the unknown function  $F$  must involve only the momentum of the spectator quark #3. A part of the  $\eta_3, \eta_3'$  dependence has been absorbed in the  $\hat{G}$  function on the right, so as to satisfy the requirements of  $S_3$ -symmetry for this 3D quantity [53].

As to the remaining factor  $F$ , it is necessary to choose its form in a careful manner so as to conform to the conservation of 4-momentum for the *free* propagation of the spectator between two neighbouring vertices, consistently with the symmetry between  $p_3$  and  $p_3'$ . A possible choice consistent with these conditions is:

$$F(p_3, p_3') = C_3 \Delta_3^{-1} \delta(\nu_3 - \nu_3') \quad (6.3.4)$$

Here  $\Delta_3^{-1}$  represents the “free” propagation of quark #3 between successive vertices, while  $C_3$  represents some residual effects which may at most depend on the 3-momentum  $\hat{p}_3$ , but must satisfy the main constraint that the 3D BSE, (6.2.12), be *explicitly* satisfied.

To check the self-consistency of the ansatz (6.3.4), integrate both sides of (6.3.2) w.r.t.  $ds_3 ds_3' dt_3 dt_3'$  to recover the 3D  $S_3$ -invariant  $\hat{G}$ -function on the left hand side. Next, in the first form on the right hand side, integrate w.r.t.  $ds_3 ds_3'$  on the  $G$ -function which alone involves these variables. This yields the quantity  $\tilde{G}_{3\xi}$ . At this stage, employ the ansatz (6.3.4) to integrate over  $dt_3 dt_3'$ . Consistency with the 3D BSE, eq.(6.2.12), now demands

$$C_3 \int \int d\nu_3 d\nu_3' \Delta_3^{-1} \delta(\nu_3 - \nu_3') = 1; (\text{since } dt = d\nu) \quad (6.3.5)$$

The 1D integration w.r.t.  $d\nu_3$  may be evaluated as a contour integral over the propagator  $\Delta^{-1}$ , which gives the pole at  $\nu_3 = \hat{\omega}_3/M$ , (see below for its definition). Evaluating the residue then gives

$$C_3 = i\pi/(M\hat{\omega}_3); \quad \hat{\omega}_3^2 = m_q^2 + \hat{p}_3^2 \quad (6.3.6)$$

which will reproduce the 3D BSE, eq.(6.2.12), *exactly*! Substitution of (6.3.4) in the second form of (6.3.2) finally gives the desired 3-body generalization of (6.1.12) in the form

$$3G(\xi\eta; \xi'\eta') = \sum_{123} D(\hat{q}_{12}) \Delta_{1F} \Delta_{2F} D(\hat{q}'_{12}) \Delta_{1F}' \Delta_{2F}' \hat{G}(\hat{\xi}_3\hat{\eta}_3; \hat{\xi}_3'\hat{\eta}_3') [\Delta_{3F}/(M\pi\hat{\omega}_3)] \quad (6.3.7)$$

where for each index,  $\Delta_F = -i\Delta^{-1}$  is the Feynman propagator. To find the effect of the ansatz (6.3.4) on the 4D BS wave function  $\Phi(\xi\eta; P)$ , we do a spectral reduction like (6.2.13) for the 4D Green's function  $G$  on the LHS of (6.3.2). Equating the residues on both sides gives the desired 4D-3D connection between  $\Phi$  and  $\phi$ :

$$\Phi(\xi\eta; P) = \sum_{123} D(\hat{q}_{12}) \Delta_1^{-1} \Delta_2^{-1} \phi(\hat{\xi}\hat{\eta}; P) \times \sqrt{\frac{\delta(\nu_3 - \hat{\omega}_3/M)}{M\hat{\omega}_3\Delta_3}} \quad (6.3.8)$$

defines the 4D wave fn in terms of piecewise vertex fns  $V_i$ , as

$$\Phi(p_1 p_2 p_3) \equiv \frac{V_1 + V_2 + V_3}{\Delta_1 \Delta_2 \Delta_3} \quad (6.3.9)$$

From (6.3.8-9), we infer the baryon- $qqq$  vertex function  $V_3$  corresponding to the ‘last’ interaction in the 12-pair as

$$V_3 = D(\hat{q}_{12}) \phi(\hat{\xi}, \hat{\eta}) \times \sqrt{2\Delta_3 \delta(\nu_3^2 M^2 - \hat{\omega}_3^2)} \quad (6.3.10)$$

and so on cyclically. (The argument of the  $\delta$ -function inside the radical for  $V_3$  simplifies to  $p_3^2 + m_q^2$ ). This expression had been obtained earlier from intuitive considerations [54b].

To account for the appearance of the 1D  $\delta$ -fn under radical in (6.3.10), it is explained elsewhere [53] that it has nothing to do with connectedness [58] as such, but merely reflects a ‘dimensional mismatch’ due to the 3D nature of the pairwise kernel  $K$  [24] imbedded in a 4D Hilbert space. (For a physical explanation, see [53]). A further self-consistency check on (6.3.10), is found by taking the limit of a point interaction, which amounts to setting  $K = \text{Constant}$ , when the radical (expectedly) disappears, and gives a Lorentz-invariant result [53], in agreement with the so-called NJL-Faddeev (contact) model [59] for 3-particle scattering. For the fermion  $qqq$  case with pairwise gluonic interactions, the details may be found in [60], wherein the strength of the ‘color’  $qq$  interaction [29] is half of that of  $q\bar{q}$  [28]. For brevity, we skip the MYTP [26] derivation of the 4D  $qqq$  vertex function under Cov LF [37] conditions, which parallels that for the 2-body case [Sect.5], except for the remark that the old-fashioned LF/NP treatment [38] gives the same results as the more formal Cov LF treatment in Sect.5, so that a similar Cov LF form for  $qqq$  dynamics should be expected [38], with  $D(\hat{q}) \rightarrow D_n(\hat{q})$ , etc in (6.3.10).

## 7 Triangle Loops Under MYTP On Cov LF/NP

In this Section, we shall illustrate the MYTP techniques on the covariant light-front to bring out the main feature, viz., structure of the triangle loop integrals free from the anomalies of time-like momenta in the product of gaussian vertex functions, such as complexities in the pion form factor [36] (see Sects. 1 and 5). To that end, we shall mainly consider the mathematical structure of the P-meson form factor, followed by a brief sketch of the structure of 3-hadron form factors, in the next few subsections, leaving routine calculational details to [37,32].

### 7.1 Pion Form Factor by Cov LF/NP Method

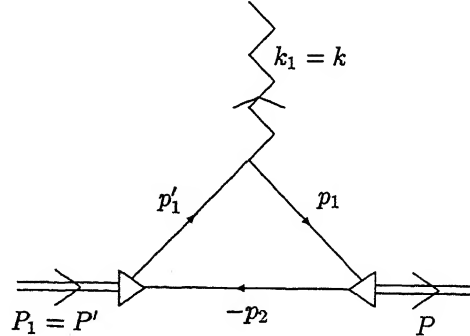


Figure 1: Triangle loop for e.m. vertex

Using fig.1 above, and an identical one with  $1 \rightarrow 2$ , (c.f., figs. 1a,1b of [34b]), the Feynman amplitude for the  $h \rightarrow h' + \gamma$  transition is given by [34b]

$$2\bar{P}_\mu F(\mathbf{k}^2) = 4(2\pi)^4 N_n(P) N_n(P') e\hat{m}_1 \int d^4 T_\mu^{(1)} \frac{D_n(\hat{q})\phi(\hat{q})D_n(\hat{q}')\phi(\hat{q}')}{\Delta_1\Delta'_1\Delta_2} + [1 \Rightarrow 2]; \quad (7.1)$$

$$4T_\mu^{(1)} = \text{Tr}[\gamma_5(m_1 - i\gamma.p_1)i\gamma_\mu(m_1 - i\gamma.p'_1)\gamma_5(m_2 + i\gamma.p_2)]; \quad \Delta_i = m_i^2 + p_i^2; \quad (7.2)$$

$$p_{1,2} = \hat{m}_{1,2}P \pm q; \quad p'_{1,2} = \hat{m}_{1,2}P' \pm q' \quad p_2 = p'_2; \quad P - P' = p_1 - p'_1 = k; \quad 2\bar{P} = P + P'. \quad (7.3)$$

After evaluating the traces and simplifying,  $T_\mu$  becomes

$$T_\mu^{(1)} = (p_{2\mu} - \bar{P}_\mu)[\delta m^2 - M^2 - \Delta_2] - k^2 p_{2\mu}/2 + (\Delta_1 - \Delta'_1)k_\mu/4 \quad (7.4)$$

The last term in (7.4) is non-gauge invariant, but it does not survive the integration in (7.1), since the coefficient of  $k_\mu$ , viz.,  $\Delta_1 - \Delta'_1$  is antisymmetric in  $p_1$  and  $p'_1$ , while the rest of the integrand in (7.1) is symmetric in these two variables. Next, to bring out the proportionality of the integral (7.1) to  $\bar{P}_\mu$ , it is necessary to resolve  $p_2$  into the mutually perpendicular components  $p_{2\perp}$ ,  $(p_2 \cdot k/k^2)k$  and  $(p_2 \cdot \bar{P}/\bar{P}^2)\bar{P}$ , of which the first two will again not survive the integration, the first due to the angular integration, and the second due to the antisymmetry of  $k = p_1 - p'_1$  in  $p_1$  and  $p'_1$ , just as in the last term of (7.4). The third term is explicitly proportional to  $\bar{P}_\mu$ , and is of course gauge invariant since  $\bar{P} \cdot k = 0$ . (This fact had been anticipated while writing the LHS of (7.4)). Now with the help of the results

$$p_2 \cdot \bar{P} = -\hat{m}_2 M^2 - \Delta_1/4 - \Delta'_1/4; \quad 2\hat{m}_2 = 1 - (m_1^2 - m_2^2)/M^2; \quad \bar{P}^2 = -M^2 - k^2/4, \quad (7.5)$$

it is a simple matter to integrate (7.1), on the lines of Sec.5, noting that terms proportional to  $\Delta_1 \Delta_2$  and  $\Delta'_1 \Delta_2$  will give zero, while the non-vanishing terms will get contributions only from the residues of the  $\Delta_2$ -pole, eq.(5.15). Before collecting the various pieces, note that the 3D gaussian wave functions  $\phi, \phi'$ , as well as the 3D denominator functions  $D_n, D'_n$ , do *not* depend on the time-like components  $p_{2n}$ , so that no further pole contributions accrue from these sources. (It is this problem of time-like components of the internal 4-momenta inside the gaussian  $\phi$ -functions under the CIA approach [24], that had plagued a earlier CIA study of triangle diagrams [36]). To proceed further, it is now convenient to define the quantity  $\bar{q} \cdot n = p_2 \cdot n - \hat{m}_2 \bar{P} \cdot n$  to simplify the  $\phi$ - and  $D_n$ -functions. To that end define the symbols:

$$(q, q') = \bar{q} \pm \hat{m}_2 k/2; \quad z_2 = \bar{q} \cdot n / \bar{P} \cdot n; \quad \hat{k} = k \cdot n / \bar{P} \cdot n; \quad (\theta_k, \eta_k) = 1 \pm \hat{k}^2/4 \quad (7.6)$$

and note the following results of pole integration w.r.t.  $p_{2n}$  [38]:

$$\int dp_{2n} \frac{1}{\Delta_2} [1/\Delta_1; 1/\Delta'_1; 1/(\Delta_1 \Delta'_1)] = [1/D_n; 1/D'_n; 2p_2 \cdot n / (D_n D'_n)] \quad (7.7)$$

Details of further calculation of the form factor are given in [37]. An essential result is the normalizer  $N_n(P)$  of the hadron, obtained by setting  $k_\mu = 0$ , and demanding that  $F(0) = 1$ . The reduced (Lorentz-invariant) normalizer  $N_H = N_n(P)P \cdot n/M$  is given by [32,37]:

$$N_H^{-2} = 2M(2\pi)^3 \int d^3 \hat{q} e^{-\hat{q}^2/\beta^2} [(1 + \delta m^2/M^2)(\hat{q}^2 - \lambda/4M^2) + 2\hat{m}_1 \hat{m}_2 (M^2 - \delta m^2)] \quad (7.8)$$

where the internal momentum  $\hat{q} = (q_\perp, M z_2)$  is formally a 3-vector, in conformity with the 'angular condition' [21]. The corresponding expression for the form factor is [32, 37]:

$$F(k^2) = 2MN_H^2(2\pi)^3 \exp[-(M\hat{m}_2 \hat{k}/\beta)^2/4\theta_k](\pi\beta^2)^{3/2} \frac{\eta_k}{\sqrt{\theta_k}} \hat{m}_1 G(\hat{k}) + [1 \Rightarrow 2] \quad (7.9)$$

where  $G(\hat{k})$  is a function of  $\hat{k}$ ; see eqs.(A.12-13) of [32].

## 7.2 'Lorentz Completion' for $F(k^2)$

The expression (7.9) for  $F(k^2)$  still depends on the null-plane orientation  $n_\mu$  via the dimensionless quantity  $\hat{k} = k \cdot n / P \cdot n$  which while having simple Lorentz transformation properties, is nevertheless *not* Lorentz invariant by itself. To make it explicitly Lorentz invariant, we shall employ a simple method of 'Lorentz completion' which is merely an extension of the 'collinearity trick' employed at the quark level, viz.,  $P_\perp \cdot q_\perp = 0$ ; see eq.(5.11). Note that this collinearity ansatz has already

become redundant at the level of the Normalizer  $N_H$ , eq.(7.8), which owes its Lorentz invariance to the integrating out of the null-plane dependent quantity  $z_2$  in (7.8). This is of course because  $N_H$  depends only on one 4-momentum (that of a *single hadron*), so that the collinearity assumption is exactly valid. However the form factor  $F(k^2)$  depends on *two independent* 4-momenta  $P, P'$ , for which the collinearity assumption is non-trivial, since the existence of the perpendicular components cannot be wished away! Actually the quark-level assumption  $P_\perp \cdot q_\perp = 0$  has, so to say, got transferred, via the  $\hat{q}$ -integration in eq.(7.9), to the *hadron level*, as evidenced from the  $\hat{k}$ -dependence of  $F(k^2)$ ; therefore an obvious logical inference is to suppose this  $\hat{k}$ -dependence to be the result of the collinearity ansatz  $P_\perp \cdot P'_\perp = 0$  at the hadron level. Now, under the collinearity condition, one has

$$P \cdot P' = P_\perp \cdot P'_\perp + P \cdot n P' \cdot \bar{n} + P' \cdot n P \cdot \bar{n} = P \cdot n P'_n + P' \cdot n P_n; \quad P \cdot \bar{n} \equiv P_n. \quad (7.10)$$

Therefore 'Lorentz completion' (the opposite of the collinearity ansatz) merely amounts to reversing the direction of the above equation by supplying the (zero term)  $P_\perp \cdot P'_\perp$  to a 3-scalar product to render it a 4-scalar! Indeed the process is quite unique for 3-point functions such as the form factor under study, although for more involved cases (e.g., 4-point functions), further assumptions may be needed.

In the present case, the prescription of Lorentz completion is relatively simple, being already contained in eq.(7.10). Thus since  $P, P' = \bar{P} \pm k/2$ , a simple application of (7.10) gives

$$\begin{aligned} k \cdot n k_n &= +k^2; \quad \bar{P} \cdot n \bar{P}_n = -M^2 - k^2/4; \\ \hat{k}^2 &= \frac{4k^2}{4M^2 + k^2} = 4\theta_k - 4 = 4 - 4\eta_k \end{aligned} \quad (7.11)$$

This simple prescription for  $\hat{k}$  automatically ensures the 4D (Lorentz) invariance of  $F(k^2)$  at the hadron level. (For comparison with alternative methods [22b], see [37]).

### 7.3 QED Gauge Corrections to $F(k^2)$

While the 'kinematic' gauge invariance of  $F(k^2)$  has already been ensured in Sec.7.1 above, there are additional contributions to the triangle loops - figs.1a and 1b of [34b] - obtained by inserting the photon lines at each of the two vertex blobs instead of on the quark lines themselves. These terms arise from the demands of QED gauge invariance, as pointed out by Kisslinger and Li [61] in the context of two-point functions, and are simulated by inserting exponential phase integrals with the e.m. currents. However, this method (which works ideally for *point* interactions) is not amenable to *extended* (momentum-dependent) vertex functions, and an alternative strategy is needed, as described below.

The way to an effective QED gauge invariance lies in the simple-minded substitution  $p_i \rightarrow p_i - e_i A(x_i)$  for each 4-momentum  $p_i$  (in a mixed  $p, x$  representation) occurring in the structure of the vertex function. This amounts to replacing each  $\hat{q}_\mu$  occurring in  $\Gamma(\hat{q}) = D(\hat{q})\phi(\hat{q})$ , by  $\hat{q}_\mu - e_q \hat{A}_\mu$ , where  $e_q = \hat{m}_2 e_1 - \hat{m}_1 e_2$ , and keeping only first order terms in  $A_\mu$  after due expansion. Now the first order correction to  $\hat{q}^2$  is  $-e_q \hat{q} \cdot \hat{A} - e_q \hat{A} \cdot \hat{q}$ , which simplifies on substitution from eq.(7.11) to

$$-2e_q \hat{q} \cdot A \equiv -2e_q A_\mu [\hat{q}_\mu - \hat{q} \cdot n \bar{n}_\mu + P \cdot \bar{n} \hat{q} \cdot n n_\mu / P \cdot n] \quad (7.12)$$

The net result is a first order correction to  $\Gamma(\hat{q})$  of amount  $e_q j(\hat{q}) \cdot A$  where

$$j(\hat{q})_\mu = -4M_> \bar{q}_\mu \phi(\hat{q}) (1 - (\hat{q}^2 - \lambda/4M^2)/2\beta^2) \quad (7.13)$$

The contribution to the P-meson form factor from this hadron-quark-photon vertex (4-point) now gives the QED gauge correction to the triangle loops, in the form of a similar function  $F_1(k^2)$  which works out as [37]:

$$F_1(k^2) = 4(2\pi)^4 N_H^2 e_q \hat{m}_1 M_>^2 \int d^4 q (M^2 - \delta m^2) \phi \phi' \left[ \frac{D'_n \hat{q} \cdot P}{\Delta'_1 \Delta_2 P' \cdot n} + \frac{D_n \hat{q}' \cdot P}{\Delta_1 \Delta_2 P \cdot n} \right] + \{1 \Rightarrow 2\} \quad (7.14)$$

where  $M_> = \sup\{M; m_1 + m_2\}$  [37], and the common factor  $2\bar{P}_\mu$  has been extracted as for  $F(k^2)$  in (7.1). Note that  $e_q$  is antisymmetric in '1' and '2', signifying a change of sign when  $\{1 \Rightarrow 2\}$  is added on the RHS. The pole integration of  $F_1(k^2)$  now yields a result like (7.9) for  $F(k^2)$ ; see [37] for details.

The large and small  $k^2$  limits of  $F(k^2)$  and  $F_1(k^2)$  are on expected lines, and we summarise only the final results for completeness [37]. For large  $k^2$ , the functions  $F(k^2)$  and  $F_1(k^2)$  both yield the correct asymptotic form  $C/k^2$ , where  $C = 0.35 \text{ GeV}^2$ , to be compared with the experimental value [62a]  $0.50 \pm 0.10$ , and the (perturbative) QCD value [63]  $8\pi\alpha_s f_\pi^2 = 0.296$ .

For low  $k^2$ , on the other hand, an expansion of  $F, F_1$  in powers of  $k^2$  yields a value of the charge radius  $R$  according to  $\langle R^2 \rangle = -\nabla_k^2 (F(k^2) + F_1(k^2))$  in the  $k^2 = 0$  limit. Of the two functions, only  $F(k^2)$  contributes in this limit [37]. The numerical values for the kaon and pion radii, vis-a-vis experiment [62b], are

$$R_K = 0.63 fm \quad vs(0.53 fm); \quad R_\pi = 0.661 fm \quad vs(0.656 fm). \quad (7.15)$$

#### 7.4 Three-Hadron Couplings Via Triangle-Loops

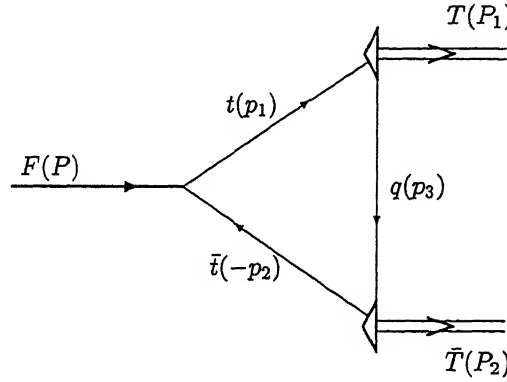


Figure 2: 3-hadron coupling

For a large class of hadronic processes like  $H \rightarrow H' + H''$  and  $H \rightarrow H' + \gamma$ , the quark triangle loop [64] represents the lowest order "tree" diagram for their evaluation. Criss-cross gluonic exchanges inside the triangle-loop are not important for this description in which the hadron-quark vertices, as well as the quark propagators are *both non-perturbative*, and thus take up a lion's share of non-perturbative effects. This is somewhat similar to the 'dynamical perturbation theory' of Pagels-Stokar [65], in which criss-cross diagrams are neglected.

We now indicate in the barest outline, the structure of the 3-hadron loop integral for the most general case of unequal mass kinematics  $m_1 \neq m_2 \neq m_3$ , while referring for notational details to ref.[32]. The full structure of the 3-hadron amplitude may be written down from fig.2 above (c.f., fig.1 of [64]), just as for the e.m. form factor (7.1):

$$A(3H) = \frac{2i}{\sqrt{3}} (2\pi)^8 \int d^4 p_i \Pi_{123} \frac{\Gamma_i(\hat{q}_i)}{\Delta_i(p_i)} \quad (7.16)$$

exhibiting cyclic symmetry, where the normalized vertex function  $\Gamma_i$  in CNPA [41] is given in an obvious notation by eq.(5.18) as

$$\Gamma_i(\hat{q}_i) = N_i(2\pi)^{-5/2} D_i(\hat{q}_i) \phi_i(\hat{q}_i); \quad D_i = 2M_i(\hat{q}_i^2 - \frac{\lambda(M_i^2, m_j^2, m_k^2)}{4M_i^2}) \quad (7.17)$$

where the 'reduced' denominator function  $D_i = D_{i+} M_i / P_{i+}$  and the (invariant) normalizer  $N_{iH}$  is  $N_i$ . The color factor and the effect of reversing the loop direction are given by  $2/\sqrt{3}$ , etc [38,64]. the overall BS normalizer [38].

To evaluate (7.16), we first write the cyclically invariant measure:

$$d^4 p_i = d^4_{\perp} \frac{1}{2} d(x_i^2) M_i^2 dy_i; \quad x_i = p_{i+}/P_{i+}; \quad y_i = p_{i-}/P_{i-} \quad (7.18)$$

The cyclic invariance of [7.18] ensures that it is enough to take any index, say 2, and first do the pole integration w.r.t. the  $y_2$  variable which has a pole at  $y_2 = \xi_2 \equiv \omega_{2\perp}^2 / (M_2^2 x_2)$ . The process can be repeated, by turn, over all the indices and the results added. Note that the  $\phi$ -functions do *not* include the time-like  $y_i$  variables under CNPA [37], so that the residues from the poles arise from only the propagators. The crucial thing to note is that the denominator functions  $D_1$  and  $D_3$  sitting at the opposite ends of the  $p_2$ -line (c.f. Fig.1 of [64]) will *cancel out* the residues from the complementary (inverse) propagators  $\Delta_3$  and  $\Delta_1$  respectively. Indeed by substituting the pole value  $y_2 = \xi_2$ , in  $\Delta_{1,3}$ , the corresponding residues in an obvious notation work out as [32]:

$$\Delta_{1,2} = \xi_2 n_{32} M_2^2 + x_2 n_{23} M_3^2 - 2\hat{\mu}_{21} M_3^2; \Delta_{3,2} = -\xi_2 n_{12} M_2^2 - x_2 n_{21} M_1^2 - 2\hat{\mu}_{23} M_1^2 \quad (7.19)$$

It is then found, with a short calculation [32], that

$$\frac{D_3(\hat{q}_3)}{\Delta_{1,2}} = 2M_3 x_2 n_{23}; \quad \frac{D_1(\hat{q}_1)}{\Delta_{3,2}} = 2M_1 x_2 n_{21} \quad (7.20)$$

which shows the precise cancellation mechanism between the  $D_i$ -functions and the residues of the propagators  $\Delta_i$  at the  $\Delta_2$  pole. This mechanism thus eliminates [24, 64] the (overlapping) Landau-Cutkowsky poles that would otherwise have caused free propagation of quarks in the loops. The same procedure is then repeated cyclically for the other two terms arising from the  $\Delta_{3,1}$  poles. Collecting the factors, the result of all the 3 contributions is compactly expressible as [64, 32]:

$$A(3H) = 8\sqrt{\frac{2\pi}{3}} \Sigma_{123} \int \int M_2 n_{23} n_{21} \pi^2 dx_2 d\xi_2 x_2^2 [TR]_2 D_2(\hat{q}_2) \Pi_{123} M_i N_i \phi_i \quad (7.21)$$

where the limits of integration for both variables are  $-\inf < (\xi_2, x_2) < +\inf$ , since these are governed, not by the on-shell dynamics of standard LF methods [22-23], but by off-shell 3D-4D BSE. The difference from [64] (under CIA [24]) arises from using CNPA [37] which has ensured that the (gaussian) functions  $\phi_i$  on the RHS of (7.21) are now free from time-like momenta (unlike in CIA [24,64]).

Eq.(7.21) is the central result of this exercise. Its general nature stems from the use of unequal mass kinematics at both the quark and hadron levels, which greatly enhances its applicability to a wide class of problems which involve 3-hadron couplings, either as complete processes by themselves (such as in decay processes) or as parts of bigger diagrams in which 3-hadron couplings serve as basic building blocks. What makes the formula particularly useful for general applications is its explicit Lorentz invariance which has been achieved through the simple method of 'Lorentz Completion' on the lines of sect.7.2 for the e.m. form factor of P-mesons; for more details, see [32].

As regards two- quark loops, such as for  $SU(2)$  mass splittings of P-mesons [33b], and the mixing of  $\rho$  and  $\omega$  off-shell propagators [33a], the distinction between CIA [24] and CNPA [37] is less sharp, (no time-like momentum problems in the overlap integrals). The same holds for one-quark loops, e.g., in the problem of vacuum condensates. For a review of these processes, as well as for other references, see [32].

## 8 Retrospect And Conclusions

In keeping with our objectives (A)-(C) defined at the outset, Sects.1-2 have attempted a panoramic view of several standard approaches to 3D BSE reductions [6-9] under the general Bethe Second Principle Philosophy of effective quark-pair interaction. In particular, the relative unfamiliarity with MYTP [26] in the literature, especially its novel feature of effecting an exact 3D reductions of the  $q\bar{q}$  and  $qqq$  BSE's, as well as exact reconstructions of the 4D amplitudes in closely parallel fashions, have necessitated the introduction of some background techniques under one roof. To that end Sect.3 collects a general derivation of the equations of motion in interlinked BSE-SDE form from an input 4-fermion Lagrangian for 'current' quarks, under MYTP conditions [25], much like the derivation of similar equations [10-12] without this constraint. And in preparation for the derivation of MYTP-governed equations in a Covariant LF/NP framework, Sect.4 collects some essential background material, especially the 'angular condition' [21], under Cov LF/NP conditions. With this background, Sect.5 outlines a comparative derivation of the MYTP-controlled 3D-4D interlinkage of  $q\bar{q}$  Bethe-Salpeter amplitudes under both CIA [24] and Cov LF [37] conditions. And in keeping with the parallelism between the 2- and 3-quark treatments, Sect.6 gives a similar derivation for the  $qqq$  system. Now this twin facility which does not seem to exist in the other 3D approaches [6-9,22-23], gives rise to a natural two-tier description [38], the 3D BSE form being appropriate for making contact with the hadron spectra [13], while the reconstructed 4D BSE yields a vertex function which allows the use of standard Feynman diagrams for 4D loop integrals. To appreciate why two distinct forms of MYTP [26] have been developed on parallel lines: Covariant Instantaneity Ansatz [24] (CIA), and Covariant Light-front [37] (Cov LF/NP), the advantage of the latter over the former in producing well-defined triangle loop integrals has been demonstrated in Sect.7 through the examples of pion form factor and more general 3-hadron couplings, except for a (less serious) problem of dependence on the 'null-plane orientation' which can be handled through a simple device of 'Lorentz completion' and yields an explicitly Lorentz-invariant structure. Similarly the baryon-quark vertex function (6.3.10) is a key ingredient of MYTP, for the calculation of baryonic loop integrals, of which the baryon self-energy [60] is the simplest example, but the calculational details [60] have been omitted for brevity.

In keeping with its mathematical (formalistic) emphasis of this Article, we have refrained from discussing the phenomenological applications, but it has been shown that the canvas of MYTP [26] is broad enough to accommodate additional physical principles. In particular, the physical basis chosen for detailed presentation, has been a QCD motivated 4-fermion Lagrangian (with an effective gluonic propagator) which generates the BSE-SDE structure by breaking its chiral symmetry dynamically ( $DB\chi S$ ) [11-12], formulated within an MYTP [26] framework.

Clearly, the **MYTP** is a very powerful Principle which helps organize a whole spectrum of phenomena under a single umbrella. For its applications, only a few examples have been indicated, but its potential warrants many more. More importantly, the interlinked 3D-4D structure of BS dynamics under **MYTP** [26] premises, gives it access to a whole range of physical phenomena, from spectroscopy to diverse types of loop integrals. The emphasis on the spectroscopy sector as an integral part of quark physics was first given by Feynman et al [39].

## References

- [1] S.Tomonaga, Prog.Theo.Phys.1,27 (1946); J.Schwinger, Phys.Rev.73, 416 (1948); R.P.Feynman, Phys.Rev.76, 749, 769 (1949); F.J.Dyson, Phys.Rev.75, 486, 1736 (1949).
- [2] I.Tamm, J.Phys.(U.S.S.R.), 9,449 (1945); S.M.Dancoff, Phys.Rev.78, 382 (1950).
- [3] H.A.Bethe and F.de Hoffmann, *Mesons and Fields, II*, Row, Peterson And Co, New York, 1955; p 199.
- [4] C.Itzykson and J.-B.Zuber, *Quantum Field Theory*, McGraw-Hill Inc, New York, 1980; Chapter 10.



- [5] E.E.Salpeter and H.A.Bethe, Phys.Rev.**84**, 1232 (1951); M.Gell-Mann and F.E.Low, *loc.cit*, 350.
- [6] (a) M. Levy, Phys.Rev.**88**, 72 (1952); (b) E.E.Salpeter, *ibid* **87**, 328 (1952).
- [7] (a) A. Logunov and A.N.Tavkhelidze, Nuovo Cimento **29**, 380 (1963); (b) V.R.Garsevanishvili, et al, Phys.Lett.**29B**, 191 (1968); (c) R.N.Faustov, Ann. Phys.(N.Y.)**78**, 176 (1973). (d) I.Todorov, Phys.Rev.**D3**, 2351 (1971)
- [8] V. Kadychevsky, Nucl.Phys.**B6**, 125 (1968);
- [9] R. Blankenbecler and R. Sugar, Phys.Rev.**142**, 105 (1966).
- [10] Review: C.D.Roberts et al, Prog.Part.Nucl.Phys.**33**, 471 (1994).
- [11] Y. Nambu and G. Jona-Lasino, Phys.Rev.**122**, 345 (1961).
- [12] (a) S.L. Adler and A.C. Davies, Nucl.Phys.**B244**, 469 (1984); (b) A.Le Yaouanc et al, Phys.Rev.**D29**, 1233 (1984); **D31**, 317 (1985); (c) R.Delbourgo and M.D.Scadron, J.Phys.G **5**, 1621 (1979).
- [13] Particle Data Group, Phys.Rev.**D54**, July 1-Part I (1996).
- [14] E.g., R.F.Meyer, Nucl.Phys.**B71**, 226 (1974).
- [15] P.A.M.Dirac, Can.J.Math.**2**, 129 (1950).
- [16] A.Komar, Phys.Rev.**D18**, 1887 (1978); L.P.Horowitz and F.Rohrlich, Phys.Rev.**D24**, 1528 (1981); H.Crater and Van Alstine, Phys.Rev.**D30**, 2585 (1984); H.Sazdjan, Phys.Lett.**156B**, 381 (1985).
- [17] (a) L.Lusanna: This Book; (b) P.P.Srivastava: This Book.
- [18] S. Weinberg, Phys.Rev.**150**, 1313 (1966)
- [19] S.Weinberg, Phys Rev.**133B**, 232 (1964).
- [20] P.A.M.Dirac, Rev.Mod.Phys.**21**, 392 (1949).
- [21] H.Leutwyler and J.Stern, Ann.Phys.(N.Y.)**112**, 94 (1978).
- [22] (a) V.A. Karmanov, Nucl.Phys.**B166**,378 (1980). (b) Review: J.Carbonell et al, Phys.Rep. (1998); to appear.
- [23] R.J.Perry,A.Harindranath and K.Wilson, Phys.Rev.Lett.**65**, 2959 (1990).
- [24] A.N. Mitra and S. Bhatnagar, Int.J.Mod.Phys.**A7**, 121 (1992).
- [25] (a) Yu. L. Kalinowski et al, Phys.Lett.**B231**, 288 (1989); (b) Yu.L. Kalinovsky et al, Few-Body Syst.**8**, (1991); (c) V.N.Pervushin et al, Fortschritte der Physik**38**, N4 (1990).
- [26] (a) M.A. Markov, Sov.J.Phys.**3**, 452 (1940); (b) H. Yukawa, Phys.Rev.**77**, 219 (1950); (c) J. Lukierski and M. Oziwicz, Phys.Lett.**B69**, 339 (1977).
- [27] (a) J.M. Cornwall et al, Phys.Rev.**D10**, 2428 (1974); (b) H. Kleinert, Phys.Lett.**B26**, 429 (1976); (c) D.W. McKay et al, Phys.Rev.**D37**, 195 (1988).
- [28] A.N. Mitra and B.M. Sodermark, Int.J.Mod.Phys.**A9**, 915 (1994).
- [29] a) A.Mittal and A.N.Mitra, Phys.Rev.Lett.**57**, 290 (1986); b) K.K. Gupta et al, Phys.Rev.**D42**, 1604 (1990); c) A. Sharma et al, Phys.Rev.**D50**, 454 (1994).
- [30] Reprint Coll: W.Buchmueller (ed), *Quarkonia*, North-Holland, 1992.

- [31] (a) M.A.Shifman et al, Nucl.Phys.**B147**, 385 (1979); (b) V.L.Chernyak and A.R.Zitnitsky, Phys.Rep.**112C**, 173 (1984); (c) B.L.Ioffe and A.V.Smigla, Nucl.Phys.**B232**, 109 (1984).
- [32] Review: A.N.Mitra, Proc.Ind.Natl.Sci.Acad.A; May-June 1999-in press.
- [33] A.N.Mitra and K.-C.Yang, Phys.Rev.**C51**, 3404 (1995); A.N.Mitra, Int J Mod Phys **A11**, 5245 (1996).
- [34] A.N.Mitra, A.Pagnamenta and N.N.Singh, Phys.Rev.Lett.**59**, 2408 (1987); N.N.Singh and A.N.Mitra, Phys.Rev.**38**, 1454 (1988).
- [35] C.R.Ji and S.Cotanch, Phys.Rev.Lett.**64**, 1484 (1990).
- [36] S.R.Chaudhury et al, Delhi Univ. Preprint (1991)–Unpublished; I.Santhanam et al, Intl.J.Mod.Phys.**E2**, 219 (1993).
- [37] A.N.Mitra, LANL hep-ph/9812404; Phys.Lett.**B463**, 293 (1999)
- [38] Review: S.Chakirabarty et al, Prog.Part.Nucl.Phys.**22**, 43-180 (1989).
- [39] R.P.Feynman, M.Kislinger and F.Ravndal, Phys.Rev.**D3**, 2706 (1971).
- [40] H.D. Politzer, Nucl.Phys.**B117**, 397 (1976).
- [41] B.A.Arbuzov et al, Mod.Phys.Lett.**A5**, 1441 (1990).
- [42] (a) GSI-ORANGE:T.Cowan et al, Phys.Rev.Lett.**56**, 444 (1986); H.Bokemeyer et al, report GSI-89-49, 1989;  
(b) GSI-EPOS:H.Tsertos et al, Z.Phys.**A326**, 2235 (1987); W.Koenig et al Phys.Lett.**B218**, 12 (1989).
- [43] J.Von Neumann and E.P.Wigner, Z.Phys.**30**, 365 (1929).
- [44] (a) R.H.Dalitz, Proc. XIII Intl Conf. on HEP, Berkeley 1966; (b) O.W.Greenberg, Phys.Rev.Lett.**13**, 564 (1964); (c) A.N.Mitra and R.Majumdar, Phys.Rev.**150**, 1194 (1966).
- [45] (a) C.Becchi and G.Morpurgo, Phys.Rev.**149**, 1284 (1966); (b) A.N.Mitra and M.H.Ross, Phys.Rev.**158**, 1630 (1967).
- [46] (a) M.Bohm, H.Joos and M.Krammer, Nucl.Phys.**B50**, 397 (1973). (b) Y.S.Kim and M.Noziere, Phys.Rev.**D8**, 3521 (1973)
- [47] (a) J.D.Bjorken, Phys.Rev.**179**, 1547 (1969); (b) R.P.Feynman, Phys.Rev.Lett.**23**, 1415 (1969).
- [48] L.Susskind, Phys.Rev.**165**, 2537 (1968).
- [49] J.Kogut and D.Soper, Phys.Rev.**D1**, 2901 (1970); J.D.Bjorken et al, Phys.Rev.**D3**, 1382 (1971).
- [50] (a) K.Fujimura, T.Kobayashi and M.Namiki, Prog.Theo.Phys.**44**,193 (1970); (b) A.Licht and A.Pagnamenta, Phys.Rev.**D2**, 1150 (1970); (c) A.Le Yaouanc et al, Phys.Rev.**D12**, 2137 (1975); (d) A.N.Mitra and I.Kumari, Phys.Rev.**D15**, 261 (1977).
- [51] S.Brodsky and G.Farrar, Phys.Rev.**D11**, 1309 (1975); V.A.Matveev et al, Lett.Nuovo. Cim.**7**, 712 (1973).
- [52] A.J.McFarlane, Rev.Mod.Phys.**34**, 41 (1981).
- [53] A.N.Mitra, LANL hep-th/9803062; Intl.J.Mod.Phys.**A14**, 4781 (1999).

- [54] (a) A.N.Mitra, *Zeits.f.Phys.C8*, 25 (1981); (b) A.N.Mitra and I.Santhanam, *Few-Body Syst.***12**, 41 (1992); (c) R. Barbieri and E.Rimiddi, *Nucl.Phys.B***141**, 413 (1978); (d) G.P.Lepage, SLAC-Preprint no.212 (1978).
- [55] (a) M.Verde, *Handbuch der Physik* **39**, 170 (1957); (b) A.N.Mitra and M.H.Ross, *Phys.Rev.***158**, 1670 (1967); (c) Yu.A.Simonov, *Sov.J.Nucl.phys.***3**, 461 (1966); (d) G.Karl and E.Obryk, *Nucl.Phys.B***8**, 609 (1968).
- [56] D.B.Liochtenberg, *Phys.Rev.***178**, 2197 (1969); A.N.Mitra and Anju Sharma, *Fortschr.Phys.***45**, 411-434 (1997).
- [57] (a) H.Kramer and M.Moshinsky, *Nucl.Phys.***82**, 241 (1966); (b) A.N.Mitra et al, *Few-Body Syst.***19**, 1 (1995); (c) J.Bijtebier, *Nuovo Cimento* **81A**, 423 (1985).
- [58] (a) L.D.Faddeev, *Sov.Phys.JETP* **12**, 1014 (1961); (b) A.N.Mitra, *Nucl.Phys.***32**, 529 (1962); (c) C.Lovelace, *Phys.Rev.***B135**, 1225 (1964); (d) S.Weinberg, *Phys.Rev.***B133**, 232 (1964).
- [59] S.Huang and J.Tjon, *Phys.Rev.***C49**, 1702 (1994); N.Ishii et al, *Austr J Phys* **50**, 123 (1997); W.Bentz, *J. Korean Phys. Soc.***29** Suppl, 5352 (1996).
- [60] A.Sharma and A.N.Mitra, LANL hep-ph/9906288; *Intl.J.Mod.Phys.***A14**, 4589 (1999).
- [61] L.S.Kisslinger and Z.Li, *Phys.Rev.Lett.***74**, 2168 (1995).
- [62] (a) E.B.Daley et al, *Phys.Rev.Lett.***45**, 232 (1980); (b) C.Bebek et al, *Phys.Rev.***D17**, 1793 (1978).
- [63] G.Farrar and D.Jackson, *Phys.Rev.Lett.***43**, 246 (1979).
- [64] W.Y.P.Hwang and A.N.Mitra, *Few-Body Syst.***15**, 1 (1993)
- [65] H.Pagels and S.Stokar, *Phys.Rev.***D20**, 2947 (1979).

# 32. The harmonic oscillator in quantum theory: A powerful bridge in physics

Marcos Moshinsky \*

Instituto de Física-UNAM.

Apartado Postal 20-364, 01000 México, D. F. México

## Abstract

An overview is given of the extensive applicational potential of the Harmonic Oscillator framework in tackling diverse problems in quantum theory, taking as basis the Author's recent book with Yu F. Smirnov) "The harmonic oscillator in modern physics" (Harwood Academic Publishers), which provides the necessary background for this Article.

## 1. Introduction

The pivotal role of the Harmonic Oscillator (HO) as a basic tool for the development of Physics, especially since the birth of quantum mechanics, hardly needs any elaboration. As the very first example of application of quantization rules, successively through spectra, wave functions, symmetries, and so on, it has had numerous applications not only in direct calculations, but also as a model for increasing our understanding of more complex problems. Indeed it pervades all dimensions of Field Theory in terms of both formulation and methodology. Its basic logic stems from the *linearity* of the field equations, or equivalently the quadratic structure of the Lagrangian/Hamiltonian, in the concerned field variable ( $\psi$ ,  $\phi$ ) which, at the level of first quantization, corresponds precisely to the quantum mechanical wave function, satisfying a (linear) Schroedinger-like equation. Indeed the language of creation/annihilation operators in QFT is based on the very structure of the harmonic oscillator which contains the co-ordinate/momentum (equivalently the field/conjugate variables in QFT) in equal proportions. For non-linear field equations too, the HO serves as a useful zero order basis, just as any quantum mechanical potential can be expanded in an HO basis. Finally, the HO forms the basis for Path Integral formulations, at the quantum mechanical as well as the field theoretic levels.

The present Article aims to draw attention to the vast applicational potential of this unique armour in the arsenal of physics through an illustrative list of topics selected from a recent book entitled *Harmonic Oscillator In Modern Physics*, (with Yu A Smirnov), which deals comprehensively with the subject. To that end, **Sect.2** is a table of contents of the various topics addressed in the book, (which is listed sequentially as ref.[24] in this Article), while **Sect.3** is a brief sketch of the subject-matter actually covered therein, without however going into their details. In **Sect.4** some recent work of the author and collaborators on relativistic particles of arbitrary spin in an harmonic oscillator potential, is summarized as a further illustration of the applicational potential of harmonic oscillator techniques.

---

\*Member of El Colegio Nacional; Email: moshi@fenix.ifisicacu.unam.mx

## 2. Table of Contents of : *The Harmonic Oscillator in Modern Physics*

### I. THE ONE-BODY PROBLEM

1. The radial wave function of the harmonic oscillator.
2. The matrix elements of  $f(r)Y_{l\mu}(\theta, \varphi)\lambda_{\mu}(\theta, \varphi$  with respect to harmonic-oscillator states. The coefficients  $B(n'l', nl, p)$ .
3. The one-electron atomic problem. Variational analysis of the ground state of the hydrogen atom in terms of harmonic- oscillator states.
4. The one-electron molecular problem. The ground state of  $H_2^+$  in terms of harmonic-oscillator states,
5. Scattering of electrons by hydrogen atoms and the form factor of the electron charge distribution in the ground state.
6. Theoretical form factor of the hydrogen atom using harmonic- oscillator states.
7. Direct determination of the ground state through a least- squares approach to the form factor. The Pseudo-Hartree-Fock (PHF) approximation.
8. The one-body harmonic-oscillator states expressed in terms of creation operators.
9. Normalization coefficients of the harmonic-oscillator states.

### II. THE TWO-BODY PROBLEM

10. Transformation brackets for two-particle harmonic-oscillator states.
11. Applications of the transformation brackets to atomic problems. The helium atom.
12. Applications of the transformation brackets to molecular problems. The  $H_2$  and  $H_3^+$  molecules.
13. Matrix elements in  $j - j$  coupling.
14. Application to two-particle problems in the 2s-1d nuclear shell. The  $^{18}O$  and  $^{18}F$  nuclei.
15. Transformation brackets for arbitrary angles in terms of standard transformation brackets.

## III. THE THREE-BODY PROBLEM

16. Matrix elements of the Hamiltonian with respect to translationally invariant states.
17. Translationally invariant three-particle states of definite permutational symmetry.
18. The general three-body problem. Applications to the lithium atom.
19. Form factors of nuclei.
20. Structure and form factor of  ${}^3\text{H}$  and  ${}^3\text{He}$ . a) Binding energy of  ${}^3\text{H}$ . b) Form factors of  ${}^3\text{H}$  and  ${}^3\text{He}$ .

## IV. THE FOUR-BODY PROBLEM

21. Harmonic-oscillator states in the symmetrical system of relative coordinates.
22. Transformation brackets between the states in the symmetrical and Jacobi coordinate system.
23. Form factor for a linear combination of harmonic-oscillator states. Application to the  $\alpha$  particle.

V. THE  $n$ -BODY PROBLEM IN THE HARTREE-FOCK APPROXIMATION

24. How good is the Hartree-Fock approximation? A simple model.
25. The set of algebraic equations and their self-consistent solution.
26. Hartree-Fock approximation with harmonic-oscillator states. The case of closed shells.
27. Applications to the beryllium atom and the  ${}^{16}\text{O}$  nucleus.

## VI. THE HARMONIC OSCILLATOR IN SCATTERING AND REACTION THEORY

28. Scattering of a particle in a central potential.
29. Solution of the equation of free motion in the harmonic oscillator basis.
30. Calculation of phase shifts.
31. Exactly soluble example.

## 32. Phase shifts and resonances in the harmonic oscillator representation

- a) s-scattering on a Gaussian potential.
- b) s-scattering on a  $\delta$ -shell potential.
- c) Evidence of resonances in variational calculations in harmonic oscillator basis.

## VII. GROUP THEORY OF HARMONIC OSCILLATORS

## 33. Group theory of the one dimensional harmonic oscillator.

## 34. The Lie algebra of linear canonical transformations.

## 35. The representation in quantum mechanics of the group of linear canonical transformations.

36. Group theory of the  $n$ -dimensional harmonic oscillator.37. Lie algebra for systems of  $m$  oscillators of  $n$  dimensions.38. The group  $U(3)$  for systems of 1 and 2 particles, as well as for 3 but only in relative motion.

- a) States for the chain (38.1).
- b) States for the chain (38.3).

39. Application of the  $U(3)$  states in the 2s-1d nuclear shell.

## VIII. FOUR DIMENSIONAL HARMONIC OSCILLATOR AND THE COULOMB PROBLEM

## 40. Eigenvalues and eigenfunctions of the four dimensional harmonic oscillator.

## 41. Kustannheim-Stiefel transformation from the four dimensional oscillator to the Coulomb problem.

## 42. Dynamic and invariance algebras of the two problems and their relations.

## IX. THE FIVE DIMENSIONAL OSCILLATOR AND NUCLEAR COLLECTIVE MOTIONS

## 43. The quadrupole liquid drop model and its classical Hamiltonian for small deformations.

## 44. The classical and quantum Hamiltonians for quadrupole deformations in the frame of reference fixed in the body.

45. States of the five dimensional oscillator characterized by irreducible representations of the chain of groups  $U(5) \supset O(3)$ .46. States associated with the chain  $U(5) \supset O(5) \supset O(3)$  in terms of traceless boson operators.

47. The expression of the states  $|klL\rangle$ , when  $L$  is even, in terms of the polynomials in the epd's of creation operators  $\eta_m$  and of the  $\alpha_m$ .
48. The polynomials  $P_{klL}(\alpha_m)$  as function of the variables  $\varphi_1, \varphi_2, \varphi_3, \gamma, \beta$  and the determination of  $\phi_K^{klL}(\gamma)$ .
49. Operators for Hamiltonians and transition probabilities for the collective model of the nucleus and their matrix elements.
50. Application to quadrupole collective motions in nuclei.
  - a) Potential energy surfaces associated with  $V(\alpha_m)$ .
  - b) Reduction of the matrix elements of the kinetic energy type to those discussed in (45.15).
  - c) Application to energy levels and transition probabilities in  $^{238}\text{U}$ .
51. Extension of the analysis to odd angular momenta.

## X. THE SIX DIMENSIONAL OSCILLATOR AND THE INTERACTING BOSON MODEL

52. The interacting boson model
53. Chains of groups associated with the six dimensional oscillator.
  - a) The chain  $U(6) \supset U(5) \supset O(5) \supset O(3)$ .
  - b) The chain  $U(6) \supset O(6) \supset O(5) \supset O(3)$ .
  - c) The chain  $U(6) \supset U(3) \supset O(3)$ .
54. Transformation brackets between the states in the  $O(6)$  and  $U(5)$  chain of groups.
55. Matrix elements of  $Q_m$  and  $Q^2$  with respect to the BIR of a  $U(5)$  chain of groups.
56. The general Hamiltonian of the interacting boson model.
57. Applications of the interacting boson model to problems of nuclear structure.

## XI. THE ONE BODY RELATIVISTIC OSCILLATOR

58. The Klein-Gordon equation with an oscillator interaction.
59. The Dirac oscillator
  - a) The symmetry Lie algebra of the Dirac oscillator.
60. The spinorial relativistic particle in an electromagnetic field.

## XII. THE TWO-BODY RELATIVISTIC OSCILLATOR

61. The system of two non-interacting spinorial particles and the relativistic cockroach nest.



62. The two-body system with a Dirac oscillator interaction. a) Solution of Eq. (62.4) when  $B = \beta_1 \beta_2$ .  
 b) Solution of Eq. (62.4) when  $B = \beta_1 \beta_2 \gamma_{51} \gamma_{52}$ .  
 c) Poincaré invariant form of the two particle system with a Dirac oscillator interaction.  
 d) Symmetry algebras and superalgebras of the two-body system with a Dirac oscillator interaction.
63. Mass spectra of the particle-antiparticle system with a Dirac oscillator interaction.  
 a) The Dirac oscillator equation for an antiparticle  
 b) Equation for the particle-antiparticle system with a Dirac oscillator interaction and its perturbative solution  
 c) Square of the mass spectra of the particle-antiparticle system  
 d) Comparison with meson spectra
64. Radial equation for the particle-antiparticle system and a qualitative application to mesons.  
 a) The radial wave equation.
65. A relativistic two body problem with an interaction modulated by a constant tensor.
66. Formulation of the oscillator problem through two independent, but constrained, relativistic equation.  
 a) The case of two scalar particles.  
 b) The case of a scalar and spinorial particle.  
 c) The case of two spinorial particles.

### XIII. THE n-BODY RELATIVISTIC OSCILLATOR

67. The non-relativistic problem for n-free scalar particles and its extension to a system with harmonic oscillator interactions.
68. The system of three relativistic scalar particles with oscillator interactions.  
 a) The spectra of the problem.  
 b) The wave function of the problem.  
 c) Poincaré invariance of the scalar three-body relativistic equation with oscillator interactions.
69. The system of three relativistic spinorial particles with a Dirac oscillator interaction.  
 a) Symmetries of the system of three quarks.
70. Spectra of the three quark system with Dirac oscillator interactions and application to the masses of non-strange baryons.  
 a) The Dirac oscillator case b) An alternative approach to the spinorial relativistic three body problem.

### 3. Summary of *The Harmonic Oscillator in Modern Physics*

Our objective in this section is to give a brief outline of the main parts of this book and to indicate the previous knowledge required to understand each of them as well as to how to use the book more effectively.

This book has 13 Chapters with 70 sections, and is divided essentially into five parts.

The first one (Chapters I to V), deals with applications of many body states with oscillator interactions, starting with one and ending with four particles, where the analysis is complete, and continuing with  $n$ -particles but only in the Hartree-Fock approximation.

To understand these first five Chapters what is required is a standard course in quantum mechanics combined with ample knowledge of angular momentum theory. For the latter, the usual Clebsch-Gordan and Racah coefficients of the rotation group have been put in the  $3j$  and  $6j$  form, at present more familiar notations, though occasionally this adds a few phase factors to the formulas. As for the permutation group the analysis is self contained except for Eq. (17.15) which is derived in standard books of group theory, though it also could be accepted as basis for Eq.(17.16-18) and the reasoning continues to be self contained from there.

The second part of the book (Chapter VI) deals with the application of harmonic oscillator states to scattering problems where, at first sight, one would think that they play no role as they vanish at infinity. Approximate phase shifts for potential scattering are obtained explicitly as well variational procedures for using harmonic oscillator states in determining resonant levels. A standard knowledge of scattering theory given in a quantum mechanics course is required.

In the third part (Chapter VII) a serious attempt is made to understand the group theory underlying the harmonic oscillator, starting with the simple case of the oscillator for one particle in one dimension and ending with  $m$  particles in  $n$  dimensions. The particular case of interest when  $n = 3$  is discussed with reference to some applications to nuclear structure in the 2s-1d shell. We tried to make the analysis completely self-contained.

The fourth part (Chapters VIII, IX, X) deals respectively with four, five, and six dimensional oscillators and their application to the Coulomb problem, the Bohr-Mottelson collective nuclear model and the Interacting Boson Model (IBM). All what is required for their understanding is given in the previous sections and in particular in Chapter VII, with the exception of section 42 where some more advanced group theoretical notions are required.

Finally the fifth part (Chapters XI, XII, XIII) deals with the relativistic many body problems with oscillator interactions, though the discussion is mainly restricted to systems of one, two and three particles and applied to the mass spectra of mesons (quark-antiquark systems) and baryons (three quark systems). Knowledge of the elements of the special theory of relativity and, in quantum mechanics, of the Dirac and Klein-Gordon equations, is assumed, though otherwise this part is again self contained.

In the conclusion we stress that we have touched only on some aspects of the harmonic oscillator in modern physics related to our own work, or to that of those with which we have come in personal contact.

### 4. The Relativistic Particle of Arbitrary Spin In A Harmonic Oscillator Potential

In this section we briefly sketch an extension of the Harmonic Oscillator framework described in the foregoing, to particles of arbitrary spins which may be characterized by the chain of groups  $SU(4) \supset [S \hat{U}(2) \otimes S \hat{U}(2)]$ . The emphasis is on the symmetry group of the problem which is the unitary symplectic subgroup  $Sp(4)$  of  $SU(4)$  (rather than of  $SU(4)$  itself). And since  $Sp(4)$  is isomorphic to  $O(5)$  we can replace it by the latter and write our equation in terms of the generators of  $O(5) \supset O(4) \supset O(3) \supset O(2)$  groups. These are more convenient as there are no multiplicity indices and the matrix elements of the generators can be given explicitly for an arbitrary irrep  $(n_1 n_2)$  of  $O(5)$ . The analysis is applied variationally to particles in an *harmonic oscillator potential* corresponding

to the irreps  $(\frac{1}{2}, \frac{1}{2}), (11), (10), (\frac{3}{2}, \frac{3}{2})$  of  $O(5)$ . We start with a brief historical introduction to the problem.

#### 4.1 Introduction

The equation of a relativistic particle of spin  $1/2$  was proposed long ago by Dirac [1] and it had an enormous success in many applications. The extension of the formalism to arbitrary spin has given rise to a veritable flood of papers in the last 50 years. Dirac himself [2] and Fierz and Pauli [3] made proposals, but which were restricted by bothersome constraints. Bargmann and Wigner [4] started not with one but a system of  $n$  Dirac type of equations and obtained a particle of spin  $n/2$  by restricting the wave function to the symmetric solution under permutation. Kemmer [5] managed to obtain a Dirac type of equation but only for spins 0 and 1. In fact Mathews [6], under the strict restrictions with which he worked argued that there could be no relativistic equations with spin higher than 1. Bhabha [7] on the other hand again returned to the possibility of arbitrary spin, though connecting them later with representations of  $SO(5)$  group as discussed by Krajcik and Nieto [8]. Weinberg [9] derived the Feynman rules for any spin in which the propagators involve matrices that transform like symmetric traceless tensors of rank  $2j$ . Nikitin [10] and his collaborators deal elegantly with relativistic particles of arbitrary spin in Coulomb and magnetic monopole fields.

In view of the above references, and possible hundreds more that seem less relevant, one could well ask if there is any reason to deal with the subject of a relativistic particle with arbitrary spin with or without interaction. There were two main reasons for getting in this crowded field. The first one was the decision to follow the Barut approach [11] that he used [12] to get a single relativistic equation for a many body problem, and particularize it to a single particle thus having only one position  $\mathbf{r}$  and momentum  $\mathbf{p}$  vectors but many  $\alpha$ 's,  $\beta$ 's in the equation. As each  $\alpha$ 's is associated with spin  $\frac{1}{2}$ ,  $n$  of them allow us to reach spins up to  $(n/2)$ . The second reason was due to noticing that  $\alpha$ 's and  $\beta$ 's could be represented by a direct product of the ordinary spin, and a new concept with the same properties as the latter which was given the name of *sign* spin.

Thus the problem becomes very similar to the one in nuclear physics in which we have ordinary spin and isospin, and the main symmetry group goes from  $SU(2)$  to  $SU(4)$ , where the latter gives rise to supermultiplets [13].

The formalism developed could be applied to any type of interaction but for simplicity we shall restrict ourselves to an harmonic oscillator potential.

Once we have the Hamiltonian operator of our equation appropriately formulated, we shall indicate a complete basis formed from the standard harmonic oscillator states in configuration space combined with the spin part, with the help of which this Hamiltonian can be transformed into an hermitian matrix of infinite number of components. Restricting ourselves to a given maximum number of quanta in our oscillator, the matrix becomes finite, thus giving us the possibility of diagonalizing it to get the energy eigenvalues for different representation of the groups involved, which in turn contain a finite number of possibilities for spin. We shall discuss qualitatively the energy spectra when  $V(r) = \frac{1}{2}m\Omega^2 r^2$ .

#### 4.2 The Hamiltonian expressed in appropriate chains of group

We use *c.g.s.* units in which we shall indicate momenta and positions by

$$\mathbf{r}', \mathbf{p}' \quad (1)$$

where we use this notations to reserve  $\mathbf{r}, \mathbf{p}$  for more appropriate units. The Dirac equation for a spin  $\frac{1}{2}$  particle in an external field can be written as

$$[c\boldsymbol{\alpha} \cdot \mathbf{p}' + mc^2\beta + V(r')]\psi = E'\psi \quad (2)$$

where

$$\boldsymbol{\alpha} = \begin{pmatrix} 0 & \boldsymbol{\sigma} \\ \boldsymbol{\sigma} & 0 \end{pmatrix}, \beta = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix} \quad (3)$$

and  $\sigma$  is the vector of the  $2 \times 2$  Pauli spin matrices.

Our next point is to note that the  $4 \times 4$  matrices  $\alpha, \beta$  in (3) can be converted into direct products of  $2 \times 2$  ones by introducing the definitions[1]

$$\hat{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, s_1 = \frac{1}{2} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, s_2 = \frac{1}{2} \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, s_3 = \frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad (4)$$

$$\check{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, t_1 = \frac{1}{2} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, t_2 = \frac{1}{2} \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, t_3 = \frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \quad (5)$$

$$\alpha = 4s \otimes t_1, \quad \beta = 2\hat{I} \otimes t_3,$$

Using (5) we can then write (2) as

$$\left\{ 4c \sum_{i=1}^3 (s_i \otimes t_i) p'_i + 2mc^2 (\hat{I} \otimes t_3) + V(r') \right\} \psi = E' \psi \quad (6)$$

where  $s_i, t_i, i = 1, 2, 3$  are the standard matrices for ordinary and sign spins given in (4) and  $\otimes$  stands for the direct product. The  $E'$  indicates the energy in *c.g.s.* units. When we want to go to a problem of larger spins[1] we introduce an index  $u = 1, 2, \dots, n$  for all the variables appearing in (6), sum the corresponding Hamiltonians and then make all  $r'_u, p'_u$  equal to a single  $r', p'$  thus getting an equation of the form

$$\sum_{u=1}^n \left\{ 4c \sum_{i=1}^3 (s_{iu} \otimes t_{iu}) p'_i + 2mc^2 (\hat{I} \otimes t_{3u}) + nV(r') \right\} \psi = nE' \psi. \quad (7)$$

whose spin can range from  $(n/2), (n/2) - 1, \dots, (1/2)$  or 0.

Now we define

$$S_i \equiv \sum_{u=1}^n (s_{iu} \otimes \check{I}), \quad R_{ij} \equiv \sum_{u=1}^n (s_{iu} \otimes t_{ju}), \quad T_i \equiv \sum_{u=1}^n (\hat{I} \otimes t_{iu}) \quad (8)$$

with  $S_i, T_i, i = 1, 2, 3$  being respectively the components of the total ordinary and sign spins, which together with the nine  $R_{ij}$ 's are the 15 generators of SU(4) group as shown by their commutation relations[14,15,16]

$$\begin{aligned} [S_i, S_j] &= i\epsilon_{ijk} S_k, \quad [T_i, T_j] = i\epsilon_{ijk} T_k, \quad [S_i, T_j] = 0, \\ [S_i, R_{jk}] &= i\epsilon_{ijl} R_{lk}, \quad [T_i, R_{jk}] = i\epsilon_{ikl} R_{jl}, \\ [R_{ij}, R_{kl}] &= \frac{1}{4} i\epsilon_{ikm} S_m \delta_{jl} + \frac{1}{4} i\delta_{ik} \epsilon_{jln} T_n. \end{aligned} \quad (9)$$

Using the definitions (8) we can rewrite Eq. (7) as

$$[4c \sum_{i=1}^3 R_{i1} p'_i + 2mc^2 T_3 + nV(r')] \psi = nE' \psi \quad (10)$$

This last equation can not, in general, be solved exactly and thus we need a convenient complete basis in which to express the operator in the square bracket in (10) as a numerical matrix.

The first requirement concerns a basis for the ordinary and sign spins. In [14] we characterized them by the chain of groups  $U(4) \supset [S \hat{U}(2) \otimes S \check{U}(2)]$ , familiar in nuclear physics[13] when we combine the ordinary spin with the isospin to get *supermultiplets*. The states can then be expressed by the kets  $|\{h\} \gamma s \sigma \tau\rangle$ , where  $\{h\} = [h_1 h_2 h_3 h_4]$  is the partition of  $n$  corresponding to

the representation of  $U(4)$ , while  $s(s+1), t(t+1)$  are the eigenvalues of Casimir operators associated with the  $SU(2)$  of ordinary and sign spins, and  $\sigma, \tau$  characterize the corresponding orthogonal  $O(2)$  subgroups of  $SU(2)$ .

We notice that the ket in the previous paragraph has an extra index  $\gamma$ , that serves to distinguish representations of  $\hat{S}\hat{U}(2) \otimes \hat{S}\hat{U}(2)$  that appear more than once in a given representation of  $SU(4)$ . This feature complicates greatly the *general* representation of  $R_{ij}$  in the basis of the ket given above. Thus we decided to follow another chain using the fact  $SU(4)$  is isomorphic to the orthogonal group  $O(6)$ .

The generators of  $O(6)$  can be characterized by the antisymmetric operators  $\Lambda_{mm'} = -\Lambda_{m'm}$  with  $m, m' = 1, 2, 3, 4, 5, 6$  and thus there are 15 of them that satisfy the commutation rules[17]

$$[\Lambda_{mm'}, \Lambda_{nn'}] = i[\delta_{m'n} \Lambda_{n'm} + \delta_{mn'} \Lambda_{nm'} + \delta_{mn} \Lambda_{m'n'} + \delta_{m'n'} \Lambda_{mn}] \quad (11)$$

Comparing them with the commutation rules Eq. (3.11) of Ref. [14] we easily see that  $\Lambda_{mm'}$  with  $m < m'$  (to avoid the repetition due to the antisymmetry) are correlated with  $S_i, R_{ij}, T_j, i, j = 1, 2, 3$  in the following way:

$$\begin{aligned} \frac{1}{2}\epsilon_{ijk}\Lambda_{jk} &= S_i \\ \Lambda_{i4} &= 2R_{i1} \\ \Lambda_{i5} &= 2R_{i2} \\ \Lambda_{i6} &= 2R_{i3} \\ \Lambda_{45} &= T_3 \\ \Lambda_{46} &= -T_2 \\ \Lambda_{56} &= T_1 \end{aligned} \quad (12)$$

where  $i, j, k$  take the values 1,2,3 and repeated indices are summed over these values.

Now  $O(6)$  has the following chain of subgroups  $O(6) \supset O(5) \supset O(4) \supset O(3) \supset O(2)$  whose generators in terms of the operators (8) are given by

$$\begin{array}{llll} 15 & S_i, R_{ij}, T_i & i, j = 1, 2, 3 & O(6) \\ 10 & S_i, R_{i1}, R_{i2}, T_3 & i = 1, 2, 3 & O(5) \\ 6 & S_i, R_{i1} & i = 1, 2, 3 & O(4) \\ 3 & S_i & i = 1, 2, 3 & O(3) \\ 1 & S_3 & & O(2) \end{array} \quad (13)$$

where on the left hand side we give the number of generators. Using (9), we easily check that the generators of each subgroup close under commutation.

We note now that in Eq. (10) only  $R_{i1}$  and  $T_3$  appear so we can restrict, ourselves to  $O(5)$  as the symmetry group. Nevertheless we consider the representations starting from  $O(6)$  as we would like to characterize our kets also by the  $\{h\}$  which is the partition of  $n$  that was mentioned above and characterizes the irreducible representation of  $U(4)$ .

To achieve the purpose of the last paragraph we note that

$$\Lambda_{12} = S_3, \quad \Lambda_{45} = T_3, \quad \Lambda_{36} = 2R_{33} \quad (14)$$

commute among themselves as seen from the relations (9) or (11). They could then be considered as three weight generators[18] of the  $O(6)$  group, while the 12 that remain can be divided into groups of 6 each, corresponding to the raising and lowering generators of the group mentioned. If we consider the state of highest weight[19], which is an eigenstate of the operators (14), the corresponding eigenvalues can be denoted by

$$q_1, q_2, q_3 \quad (15)$$

Turning now our attention to  $U(4)$  group, its generators are of the form

$$C_r^{r'}, r < r'; \quad C_r^{r'}, r = r'; \quad C_r^{r'}, r > r' \quad (16)$$

where  $r, r' = 1, 2, 3, 4$  and we indicate the raising, weight and lowering generators. Again if we have a state of highest weight it would be an eigenstate of  $C_1^1, C_2^2, C_3^3, C_4^4$  and their eigenvalues

$$h_1, h_2, h_3, h_4 \quad (17)$$

characterize the irreducible representations of  $U(4)$ . In table II of reference [16] we give, in spherical component form, the  $S_i, T_i, R_{ij}$  as linear functions of  $C_r^{r'}$  and, in particular, for the weight generators, where the index 0 is equivalent to the index 3 of cartesian components, we have that (15) are related to (17) by [16]

$$q_1 = \frac{1}{2}(h_1 + h_2 - h_3 - h_4), q_2 = \frac{1}{2}(h_1 - h_2 + h_3 - h_4), q_3 = \frac{1}{2}(h_1 - h_2 - h_3 + h_4), \quad (18)$$

As  $h_1 + h_2 + h_3 + h_4 = n$  with  $h_1 \geq h_2 \geq h_3 \geq h_4$  we see that  $q_1, q_2, q_3$  can be integer or semi integer numbers depending on whether  $n$  is even or odd and furthermore  $q_1 \geq q_2 \geq q_3$  with  $q_1, q_2$  being positive while, in some cases,  $q_3$  could also take negative values.

Having established the relation between the irreducible representations (irreps) of  $O(6)$  and  $U(4)$ , we turn our attention to  $O(5)$  which, as we indicated before, is a smaller symmetry group for the Hamiltonian in Eq.(10) as, using (12), it can be written in the form

$$[2c \sum_{i=1}^3 \wedge_{i4} p_i' + 2mc^2 \wedge_{45} + nV(r')] \psi = nE' \psi. \quad (19)$$

### 4.3 Matrix elements of the generators $\wedge_{45}, \wedge_{i4}, i = 1, 2, 3$ in a basis of irreps in the chain $O(5) \supset O(4) \supset O(3) \supset O(2)$

As is well known [20] the irreps of  $O(2k+1)$  and  $O(2k)$  are characterized by partitions involving only  $k$  numbers that can be integer or seminteger and non-negative, except for the last one in the even case which sometimes can be negative.

Rather than discussing the general theory analyzed in references [20], we shall restrict our analysis to the chain of orthogonal groups that appear in the title of this section, where the irreps will be denoted as follows:

$$\begin{aligned} O(5) ; n_1, n_2 \\ O(4) ; m_1, m_2 \\ O(3) ; s \\ O(2) ; \sigma \end{aligned} \quad (20)$$

As  $O(5)$  is a subgroups of  $O(6)$ ,  $n_1, n_2$  are restricted by the inequalities [20]

$$q_1 \geq n_1 \geq q_2 \geq n_2 \geq |q_3|. \quad (21)$$

Turning now our attention to  $O(4)$ ,  $m_1, m_2$  are restricted by the inequalities [20]

$$n_1 \geq m_1 \geq n_2 \geq |m_2|. \quad (22)$$

For  $O(3)$  we have the single number  $s$  restricted by [20,21]

$$m_1 \geq s \geq |m_2|. \quad (23)$$

Finally  $\sigma$  of  $O(2)$  is restricted by  $|\sigma| \leq s$  which implies that is given by [21]

$$\sigma = s, s-1, \dots, -s+1, -s \quad (24)$$

as all the values indicated can only change by one unit at a time within the limits indicated in the inequalities. We note then that the integer or seminteger character of the representation  $(q_1, q_2, q_3)$  of  $O(6)$  propagates to all of its subgroups.

The kets for the spin part of  $O(5) \supset O(4) \supset O(3) \supset O(2)$  chain of groups, can be denoted by

$$\left| \begin{array}{c} n_1 n_2 \\ m_1 m_2 \\ s \\ \sigma \end{array} \right\rangle \quad (25)$$

and the matrix elements of  $\Lambda_{45}, \Lambda_{34}$  with respect to them have been calculated in references [22,23]. Before giving them explicitly here, we note that  $\Lambda_{i4}$  is a Racah tensor of order 1 with respect to the  $O(3)$  group and, in particular,  $\Lambda_{34}$  corresponds to the component 0 of this tensor so we have by the Wigner-Eckart theorem that[21]

$$\left\langle \begin{array}{c} n_1 n_2 \\ m'_1 m'_2 \\ s' \\ \sigma' \end{array} \right| \Lambda_{34} \left| \begin{array}{c} n_1 n_2 \\ m_1 m_2 \\ s \\ \sigma \end{array} \right\rangle = \langle s\sigma, 10 | s'\sigma' \rangle \left\langle \begin{array}{c} n_1 n_2 \\ m'_1 m'_2 \\ s' \end{array} \parallel \Lambda_4 \parallel \begin{array}{c} n_1 n_2 \\ m_1 m_2 \\ s \end{array} \right\rangle, \quad (26)$$

where  $\langle \parallel \rangle$  is a standard  $O(3)$  Clebsch-Gordan coefficient. Thus for  $\Lambda_{14}, \Lambda_{24}$  we need only the reduced matrix element on the right hand side of (26), and its explicit value, together with that of  $\Lambda_{45}$ , is given below [22,23]

$$\begin{aligned} \left\langle \begin{array}{c} n_1 n_2 \\ m'_1 m'_2 \\ s \end{array} \parallel \Lambda_{45} \parallel \begin{array}{c} n_1 n_2 \\ m_1 m_2 \\ s \end{array} \right\rangle = \\ \frac{i}{2} \sqrt{\frac{(m_1 - s + 1)(m_1 + s + 2)(n_1 - m_1)(n_1 + m_1 + 3)(m_1 - n_2 + 1)(m_1 + n_2 + 2)}{(m_1 + m_2 + 1)(m_1 + m_2 + 2)(m_1 - m_2 + 1)(m_1 - m_2 + 2)}} \\ \times \delta_{m'_1, m_1+1} \delta_{m'_2, m_2} \\ - \frac{i}{2} \sqrt{\frac{(s - m_2)(s + m_2 + 1)(n_2 - m_2)(n_2 + m_2 + 1)(n_1 - m_2 + 1)(n_1 + m_2 + 2)}{(m_1 + m_2 + 2)(m_1 + m_2 + 1)(m_1 - m_2)(m_1 - m_2 + 1)}} \\ \times \delta_{m'_1, m_1} \delta_{m'_2, m_2+1} \\ + \frac{i}{2} \sqrt{\frac{(s + m_1 + 1)(m_1 - s)(n_1 - m_1 + 1)(n_1 + m_1 + 2)(m_1 - n_2)(m_1 + n_2 + 1)}{(m_1 + m_2)(m_1 + m_2 + 1)(m_1 - m_2)(m_1 - m_2 + 1)}} \\ \times \delta_{m'_1, m_1-1} \delta_{m'_2, m_2} \\ + \frac{i}{2} \sqrt{\frac{(s - m_2 + 1)(s + m_2)(n_2 - m_2 + 1)(n_2 + m_2)(n_1 - m_2 + 2)(m_2 + n_1 + 1)}{(m_1 + m_2)(m_1 + m_2 + 1)(m_1 - m_2 + 2)(m_1 - m_2 + 1)}} \\ \times \delta_{m'_1, m_1} \delta_{m'_2, m_2+1} \end{aligned} \quad (27)$$

$$\begin{aligned} \left\langle \begin{array}{c} n_1 n_2 \\ m_1 m_2 \\ s' \end{array} \parallel \Lambda_4 \parallel \begin{array}{c} n_1 n_2 \\ m_1 m_2 \\ s \end{array} \right\rangle = -i \sqrt{\frac{(m_1 - s)(m_1 + s + 2)(s - m_2 + 1)(s + m_2 + 1)}{(2s + 3)(s + 1)}} \delta_{s', s+1} \\ + \frac{(m_1 + 1)m_2}{\sqrt{s(s + 1)}} \delta_{s', s} + i \sqrt{\frac{(m_1 - s + 1)(m_1 + s + 1)(s - m_2)(s + m_2)}{(2s - 1)s}} \delta_{s', s-1} \end{aligned} \quad (28)$$

Note now that in Eq. (19) only  $\Lambda_{i4}, \Lambda_{45}$  appear, which are generators of  $O(5)$ , and thus  $(n_1, n_2)$ , that give the irrep of  $O(5)$ , are integrals of motion for the Hamiltonian operator in the square bracket of (19).

#### 4.4 The complete set of variational states and the matrix elements of our Hamiltonian with respect to them.

So far we have not mentioned that part of our state that is a function of  $\mathbf{r}'$  in our configuration space. As, in general, the Eq. (19) does not admit an exact solution we choose the simplest set of states, that of the harmonic oscillator, to carry our analysis variationally. As the frequency  $\omega$  of the oscillator is our only parameter, we can introduce it in the Hamiltonian [24] and thus consider only states of frequency 1 given by the ket

$$|Nl\mu\rangle = R_{Nl}(r')Y_{l\mu}(\theta, \varphi) \quad (29)$$

with  $Y$  being a spherical harmonic where  $l$  is the orbital angular momentum, while  $R$  is the radial part of the ket characterized by the number of quanta  $N$ .

As the total angular momentum

$$\mathbf{J} = \mathbf{L} + \mathbf{S}, \quad \mathbf{L} = \mathbf{r} \times \mathbf{p} \quad (30)$$

is obviously an integral of motion, we can write our full ket as

$$\left| N \begin{pmatrix} l, & n_1 n_2 \\ m_1 m_2 \\ s \end{pmatrix} jm \right\rangle = \sum_{\sigma\mu} \langle l\mu, s\sigma | jm \rangle |Nl\mu\rangle \begin{vmatrix} n_1 n_2 \\ m_1 m_2 \\ s \\ \sigma \end{vmatrix} \quad (31)$$

where  $\langle | \rangle$  is a Clebsch Gordon coefficient.

We now have to apply standard Racah algebra[21] to the state (31) and we get the following result

$$\begin{aligned} & \left\langle N \begin{pmatrix} l', & n_1 n_2 \\ m'_1 m'_2 \\ s' \end{pmatrix} jm \right| 2c \sum_{i=1}^3 \Lambda_{i4} p_i + 2mc^2 \Lambda_{45} + nV(r') \left| N \begin{pmatrix} l, & n_1 n_2 \\ m_1 m_2 \\ s \end{pmatrix} jm \right\rangle \\ &= 2(-1)^{l'+s-j} W(l' s s'; j) [(2l' + 1)(2s' + 1)]^{1/2} \\ & \langle N' l' || p' || N l \rangle \begin{vmatrix} n_1 n_2 \\ m'_1 m'_2 \\ s' \end{vmatrix} \left\| \Lambda_4 \right\| \begin{vmatrix} n_1 n_2 \\ m_1 m_2 \\ s \end{vmatrix} \right\rangle + 2\delta_{N'N} \delta_{l'l} \begin{vmatrix} n_1 n_2 \\ m'_1 m'_2 \\ s' \end{vmatrix} \left\| \Lambda_{45} \right\| \begin{vmatrix} n_1 n_2 \\ m_1 m_2 \\ s \end{vmatrix} \right\rangle \\ & + n\delta_{l'l} \delta_{m'm_1} \delta_{m'_2 m_2} \delta_{s's} \langle N' l' || V(r') || N l \rangle \end{aligned} \quad (32)$$

where we have assumed that  $V(r)$  is only a function of the magnitude of  $\mathbf{r}$ .

In (32) the  $W$  is a Racah coefficient and the reduced matrix elements of  $\Lambda_{i4}, \Lambda_{45}$  are given in (27, 28). To get then the full matrix representation of our Hamiltonian we only need determine the reduced matrix elements of  $p'$  and  $V(r')$  which will be discussed in the next section where we shall introduce more appropriate units.

#### 4.5 Example: The energy spectra of a relativistic harmonic oscillator corresponding to a definite irrep of $O(5)$

The equation is now (12) where

$$V(r') = \frac{1}{2} m \Omega^2 r'^2 \quad (33)$$

with  $m$  the mass and  $\Omega$  the frequency of the oscillator, all in *cgs* units.

We shall introduce dimensionless units by dividing (19) by  $\hbar\Omega$  and defining

$$E = (\hbar\Omega)^{-1}(E' - mc^2) \quad (34)$$

We now note that  $\mathbf{r}', \mathbf{p}'$  are in *c.g.s* units associated with a frequency  $\omega$  of the variational states. Instead of having  $\omega$  in the state (29) we introduce it in the Hamiltonian by replacing  $\mathbf{r}', \mathbf{p}'$  by dimensionless expressions through the relation

$$\mathbf{r}' = (\hbar/m\omega)^{1/2} \mathbf{r}, \quad \mathbf{p}' = (m\omega\hbar)^{-1/2} \mathbf{p} \quad (35)$$



The equation (19) can then be written as

$$\left\{ 2a\epsilon \sum_{i=1}^3 (\wedge_{i4} p_i) + a^2 (2 \wedge_{45} - n) + \frac{nr^2}{2\epsilon^2} \right\} \psi = nE\psi \quad (36)$$

where we subtracted the total rest energy and

$$a \equiv \left( \frac{mc^2}{\hbar\Omega} \right)^{1/2}, \quad \epsilon = (\omega/\Omega)^{1/2} \quad (37)$$

and  $\epsilon$  will be now the variational parameter against which we plot the energy  $E$ .

The matrix representation of the left hand side of Eq. (36) can be obtained in a way similar to (32) replacing  $r', p'$  by  $r, p$  so we need only to introduce in (32) the well know [24,25] reduced matrix elements

$$\begin{aligned} & \langle N'l' || p || Nl \rangle \\ &= \frac{i}{\sqrt{2}} \left\{ \left[ (N+l+3)^{1/2} \delta_{N'N+1} + (N-l)^{1/2} \delta_{N'N-1} \right] \sqrt{\frac{(l+1)}{(2l+3)}} \delta_{l'l+1} \right. \\ & \quad \left. + \left[ (N-l+2)^{1/2} \delta_{N'N+1} + (N+l+1)^{1/2} \delta_{N'N-1} \right] \sqrt{\frac{l}{(2l-1)}} \delta_{l',l-1} \right\} \\ & \langle N'l' || r^2 || Nl \rangle = -\frac{1}{2} \sqrt{(N-l)(N+l+1)} \delta_{N'N-2} \\ & \quad + (N + \frac{3}{2}) \delta_{N'N} - (1/2) \sqrt{(N-l+2)(N+l+3)} \delta_{N'N+2} \end{aligned} \quad (38)$$

From (27), (28), (32) and (38), (39) we can then write the matrix representation of the operator of the left hand side of (36), in which the irrep  $(n_1 n_2)$  of  $O(5)$  is an integral of motion together with the total angular momentum  $j$ . This implies that in (32),  $(n_1 n_2), j$  are fixed and the same in bra and ket.

In the following subsections we shall discuss simple particular cases of this matrix taking the lowest value of  $j$  allowed i.e.  $j = 0$  if  $n$  is even or  $j = \frac{1}{2}$  if  $n$  is odd.

In the title of the examples given below we have both the partitions  $\{h\}$  irrep of  $U(4)$ , and  $(n_1 n_2)$  irrep of  $O(5)$  related by (21), as well as the values of  $j$  mentioned in the previous phrase.

#### a) The case $\{h\} = 1, (n_1 n_2) = (\frac{1}{2} \frac{1}{2}), j = 1/2$

This corresponds to the ordinary Dirac equation in an oscillator potential and from the inequalities (22) to (24) we see that  $O(5)$  part (25) takes the form

$$\left| \begin{array}{c} \frac{1}{2} \frac{1}{2} \\ \frac{1}{2} m_2 \\ \frac{1}{2} \\ \sigma \end{array} \right\rangle \quad (39)$$

As  $j = \frac{1}{2}, l$  could only be 0 or 1 and this is given automatically if  $N$  is even or odd.

Thus the kets (31) could be written in the short hand notation

$$|N^{m_2}\rangle, \quad m_2 = \pm \frac{1}{2}, \quad N = 0, 1, 2, \dots \quad (40)$$

Ordering the kets by increasing values of  $N$  and, instead of  $m_2 = \pm \frac{1}{2}$ , just denoting it as  $m_2 = \pm$  the set of states, (where we suppress the ket notation), can be indicated by

$$0^+ 0^- 1^+ 1^- \dots 14^+ 14^- 15^+ \dots \quad (41)$$

where we end with an  $N^+$  to get a better fit to the positive energy levels. The energy levels (both positive and negative) can be plotted as a function of the variational parameter  $\epsilon$ , for different values of  $a$ , with its small/large values corresponding to the relativistic/ non-relativistic domains respectively. In particular, for  $a = 10$  we have the non-relativistic limit, and the positive energies are in fact close to the value  $E = (N + \frac{3}{2})$ , at least for the levels up to  $N = 10$ .

b)  $\{h\} = \{2\}, (n_1, n_2) = (1, 1); j = 0$

The present case concerns  $n = 2$  in which the spin  $s$  can take values 0 and 1. Since we consider here the lowest possible value of total spin  $j = 0$ ,  $l$  can take only two values viz 0 and 1. Thus even values of  $N$  occur with  $l = 0, s = 0$ , while odd  $N$  corresponds to the combination  $l = 1, s = 1$ . In the chain of groups  $O(5) \supset O(4) \supset O(3) \supset O(2)$ , the symmetric partition  $\{2\}$  can have the following ket

$$\left| \begin{array}{c} 11 \\ 1m_2 \\ s \\ \sigma \end{array} \right\rangle \quad \text{with} \quad m_2 = \pm 1, 0 \text{ and } s = 0, 1. \quad (42)$$

$s = 1$  can have all possible values of  $m_2$ , viz,  $\pm 1$  or 0 while  $s = 0$  can correspond to only  $m_2 = 0$ . Since all quantum numbers except  $s$  and  $m_2$  are fixed in the above ket (with  $\sigma$  not occurring explicitly in the matrix) we can denote the states in short hand notation as

$$|N^{sm_2}\rangle \quad (43)$$

which according to above arguments would give rise to following chain of states in the matrix.

$$0^{00}1^{11}1^{10}1^{1-1}2^{00}3^{11}3^{10}3^{1-1} \dots \quad (44)$$

We restrict the hamiltonian matrix in this chain of states to  $14^{00}$  whose diagonalized eigenvalues ( $E$ ) may again be plotted against the variational parameter  $\epsilon$  for different values of  $a$ . The general behavior of the curves is governed by Eq. (36). When  $\epsilon \rightarrow 0$  the kinetic energy as well as the rest-mass terms become negligible compared to the potential energy which goes to  $+\infty$ , thus  $E \rightarrow \infty$ . But as  $\epsilon \rightarrow \infty$ , the potential energy term goes to zero and the equation reduces to that of free particle whose energy spectra is given by [26]

$$E_m = \left( \frac{n-2m}{n} \right) \sqrt{p^2 + 1 - a^2}; \quad m = 0, 1, 2, \dots n \quad (45)$$

In particular for  $n = 2$ , there appear a set of curves corresponding to positive energies forming bound states, a negative continuum and a set which converges to  $-a^2$  as  $\epsilon \rightarrow \infty$ . The chain of states (44) gives rise to positive bound states occurring for odd values of  $N$ , thus the ground state does not exist for  $j = 0$ . Further, a gap of  $a^2$  (or  $mc^2$ ) exists between the three sets of levels starting from where bound states are formed. Small values of  $a$  correspond to a fully relativistic region (e.g.,  $a = 1$ ) wherein the bound states are formed at higher values of  $\epsilon$  which no longer follow the non-relativistic pattern.

c)  $\{h\} = \{11\}, (n_1, n_2) = (1, 0); j = 0$

The antisymmetric partition  $\{11\}$  in the canonical chain of orthogonal groups correspond to the following ket

$$\left| \begin{array}{c} 10 \\ m_1 0 \\ s \\ \sigma \end{array} \right\rangle \quad \text{with} \quad m_1 = 0, 1 \text{ and } s = 0, 1. \quad (46)$$

which can be written for convenience as

$$|N^{sm_1}\rangle \quad (47)$$

In the present case the chain of states can be written as

$$0^{01}0^{00}1^{11}2^{01}2^{00} \dots \quad (48)$$

which we restrict to  $15^{11}$  while calculating the energy matrix. the behaviour is similar to the previous case of  $\{2\}$  except that the present case gives rise to bound states with even values of  $N$ .

Table 1 gives the results for bound states for non-relativistic case  $a = 10$ , which appear at  $\epsilon = 1.02$  for all partitions discussed. Table 2 shows the bound states for fully relativistic region of  $a = 1$ . For  $\{1\}$ , they are formed at  $\epsilon = 2.5$  while for the other cases of  $j = 0, n = 2$ , they appear at  $\epsilon = 2.62$ .

Table 1  
Energy levels for non-relativistic case with  $a = 10, \epsilon = 1.02$

N	N+3/2	{1}	{11}	{2}	{3}
0	1.5	1.4991	1.5028		
1	2.5	2.4880		2.4868	2.4778
2	3.5	3.4806	3.4843		3.4520
3	4.5	4.4624		4.4613	4.4296
4	5.5	5.4479	5.4516		5.3911
5	6.5	6.4227		6.4217	6.3582
6	7.5	7.4013	7.4049		7.2849
7	8.5	8.3694		8.3685	
8	9.5	9.3412	9.3447		
9	10.5	10.3027		10.3018	
10	11.5	11.2679	11.2693		
11	12.5	12.2239		12.2229	
12	13.5	13.1829	13.2333		
13	14.5	14.1314		14.0705	
14	15.5	15.0299	15.0174		

d)  $\{h\} = \{3\}, (n_1, n_2) = (3/2, 3/2); j = 1/2$

The symmetric partition  $\{3\}$  in the canonical chain of groups  $O(5) \supset O(4) \supset O(3) \supset O(2)$  corresponds to the state function

$$\left| \begin{array}{c} 3/2, 3/2 \\ 3/2 m_2 \\ s \\ \sigma \end{array} \right\rangle \quad (49)$$

where  $m_2 = \pm 3/2, \pm 1/2$  for  $s = 3/2$  and  $m_2 = \pm 1/2$  for  $s = 1/2$ . Since  $(s, m_2)$  are the only variables (with  $\sigma$  not occurring in explicit form in the Hamiltonian (32)) and  $(n_1, n_2); m_1$  are fixed, we can use the short hand rotation  $N^{sm_2}$  for the above ket. Thus the matrix for the Hamiltonian (32) with  $j = 1/2$  would have the following chain of states

$$0^{1/2} 1^{1/2} 0^{1/2-1/2} 1^{1/2} 1^{1/2-1/2} 1^{3/2} 3^{3/2} 1^{3/2-1/2} 1^{3/2-1/2} 1^{3/2-3/2} \dots \quad (50)$$

Since for  $N = 0, l = 0; s = 1/2$  and  $m_2 = \pm 1/2$ , positive ground state occurring for  $s = 3/2$ , does not appear in this chain of states for  $j = 1/2$ . We limit the matrix up to  $7^{3/2-1/2}$  with the states occurring in the same order as in (50) which implies that the size of the matrix is  $43 \times 43$ . This particular choice of basis states is made in order to get the best fit for the physical positive bound states. In the present case of  $n = 3$ , positive states with  $s = 3/2$  occur for even as well as odd values of  $N$  as seen from (50) with no ground state occurring for  $N = 0$ . The present case of  $n = 3$  gives rise to energy curves which though starting from  $\infty$  at  $\epsilon = 0$ , get separated into four "floors" (see (45)) (one extra state appearing separately from the four floors is due to the choice of basis with no additional physical significance). Since it is only the uppermost energy floor which gives rise to the physical states (while the bound state in the other floors formed out of mixture with the negative energies, do not contribute to the physical states). We give in table 1, the corresponding minima  $E$  which occur at  $\epsilon = 1.02$  for  $a = 10$  and at  $\epsilon = 1.73$  for  $a = 1$ . Note that for non-relativistic limit of  $a = 10$ , the energies seem to follow the rule  $E = N + 3/2$ . The behaviour of all the floors is yet to be understood in the relativistic case.

Table 2  
Energy levels for relativistic case  $a = 1$

N	N+3/2	$\epsilon = 2.5$	$\epsilon = 2.62$		$\epsilon = 1.73$
		{1}	{11}	{2}	{3}
0	1.5	1.5025	1.5315		
1	2.5	2.0126		2.0291	2.5654
2	3.5	2.7806	2.8123		2.6308
3	4.5	3.3581		3.4386	3.5688
4	5.5	4.0391	4.1364		3.6134
5	6.5	4.7173		4.8720	4.7359
6	7.5	5.3458	5.5244		4.7450
7	8.5	6.1216		6.3544	
8	9.5	6.7085	6.9688		
9	10.5	7.5947		7.9121	
10	11.5	8.1450	8.4824		
11	12.5	9.1860		9.5993	
12	13.5	9.7019	10.121		
13	14.5	11.0286		11.5602	
14	15.5	11.5087	12.1073		

## Conclusions

We have derived a wave equation for a relativistic particle with arbitrary spin using the generators of the  $O(5) \supset O(4) \supset O(3)$  chain of groups. We did not discuss the Lorentz invariance of our equation as its initial formulation is in terms of the  $\alpha$ 's rather than the  $\gamma$ 's matrices. We shall use the latter in another [27] publication showing not only that the equations are Poincaré invariant, but also that they lead through our simple supermultiplet formulation to the Bhabha equation proposed long ago [7].

## Acknowledgement

The material of Sect.4 was taken from two papers in which the author participated:

I. Supermultiplets and relativistic problems I. The free particle with arbitrary spin in a magnetic field [28];

II Analysis of relativistic particles of arbitrary spin through different chains of groups [29].

Unfortunately certain graphs for the Energy levels versus the HO Quantum Numbers could not be reproduced in this Article due to technical problems. However the same may be found in ref [29].

The author would like to express his thanks to the collaborators in the two papers [28,29] and in particular to Dr. Anju Sharma for the calculation of the tables as well as to Mrs. Fanny Arenas for the preparation of the manuscript and diskette.

## References

- [1] P.A.M. Dirac "The principles of quantum mechanics" Fourth Edition (Oxford at the Clarendon Press 1959) pp.252-275.
- [2] P.A.M. Dirac, *Proc. Roy. Soc. A*, **155**, 447 (1936).
- [3] M. Fierz and W. Pauli, *Proc. Roy. Soc. A* **173**, 211 (1939).
- [4] V. Bargmann and E. P. Wigner, *Proc. Nat. Acad. Sci. (USA)* **34**, 211 (1948).

- [5] N. Kemmer, *Proc. Roy. Soc. A*, 73 (1939).
- [6] P.M. Mathews and B. Vijayalakshmi *J. Math. Phys.* 25, 1080 (1984).
- [7] H. J. Bhabha, *Rev. Mod. Phys.* 21, 451 (1949).
- [8] R. A. Krajcik and M. Martin Nieto *Am. J. Phys.* 45, 818 (1977).
- [9] S. Weinberg, *Phys. Rev.* 133, B1318, (1964).
- [10] W. I. Fushchich, A. G. Nikitin, W.M. Susloparow, *Il Nuovo Cimento* 87, A, 415, (1985).
- [11] A. O. Barut, S. Komy, *Fortsch. Phys.* 33, 6(1985); A. O. Barut and G. L. Strobel, *Few Body Systems* 1, 167 (1986).
- [12] M. Moshinsky, G. Loyola, C. Villegas *J. Math. Phys.* 32, 373 (1991).
- [13] E. P. Wigner, *Phys. Rev.* 51, 106 (1937).
- [14] M. Moshinsky and Yu. F. Smirnov *J. Phys. A: Math. Gen.* 29, 6027 (1996).
- [15] M. Moshinsky and J. G. Nagel, *Phys. Lett.* 5, 173 (1963).
- [16] M. Moshinsky, *J. Math. Phys.* 7, 691 (1966)
- [17] M. Moshinsky, *Group Theory and the many body problem* (Gordon and Breach, New York, 1968) pp. 36.
- [18] M.Moshinsky, loc. cit. p.14
- [19] I. M. Gelfand and M. L.Zetlin, *Dok. Akad. Nauk. USSR* 71, 147 (1950) (In Russian).
- [20] S.C. Pang and K. T. Hecht , *J. Math. Phys* 8, 1233 (1967) .
- [21] M. E. Rose, *Elementary Theory of Angular Momentum* (John Wiley and Sons, New York 1957) pp. 85-88, 115-119.
- [22] G. F.Filippov, V. I. Ovcharenko, Yu F. Smirnov *Microscopic theory of collective excitations of atomic nuclei*. Kiev, Nauka Dumka 1981 (In Russian) pp. 252-254
- [23] A. G. Nikitin and V. V. Tretynik, *J. Phys. A: Math. Gen.* 28, 1655 (1995)
- [24] M. Moshinsky and Yu. F. Smirnov, *The Harmonic Oscillator in Modern Physics*, (Harwood Academic Press, The Netherlands 1996) Eq. (10.35) in p. 35; Eq. (3.2) in p. 3.
- [25] M. Moshinsky and C. Quesne, *Ann. Phys. (N.Y.)* 148, 462 (1983).
- [26] M. Moshinsky, *Rev. Mex. Fis.* 43, 511 (1997)
- [27] M. Moshinsky, A. G. Nikitin, A. Sharma, Yu. F. Smirnov. *J. Phys. A:Math. Gen.* 31, 6045 (1998).
- [28] M.Moshinsky and Yu F Smirnov, *J. Phys. A: Math Gen* 29 6027 (1996).
- [29] M.Moshinsky, A. G. Nikitin, A. Sharma and Yu. F. Smirnov, *Rev. Mex. Fís.* 44 , Supl. 2, 1 (1998).

# Conclusion

33. Modern Perspectives On Foundations Of Quantum Mechanics by D.Home



# 33. Modern Perspectives On Foundations Of Quantum Mechanics

Dipankar Home\*

Dept. of Physics, Bose Institute, Calcutta 700009, India

## 1 Introduction

The foundational questions of quantum mechanics are usually dismissed on the ground that they stem primarily from subjective predilections, aesthetic considerations or from “classical” prejudices. However, what we are increasingly learning from studies in recent years is that such conceptual issues are not restricted to the realm of abstract philosophy alone but have hard-core physical relevance. A number of key foundational problems have become more precisely formulated and are either already amenable to experimental studies or promise to do so in near future with ingenious new ideas being explored, coupled with rapid advances in the relevant technology. Those who deplore the rift between science and philosophy which began to develop in the eighteenth century may take comfort in the blending of these two exhibited in the current research scenario related to the foundations of quantum mechanics. As Shimony [1] suggests, they will find here a vindication of the old sense of “Natural Philosophy”.

Here it may be useful to recall the background of the genesis of the general theory of relativity. The triggering element was Einstein’s realisation that the equivalence principle implied that one could not construct a theory of gravitation compatible with the principle of relativity by restricting the space - time transformations to those that belong to the Lorentz group. This was then a mere conceptual lacuna in the fabric of the existing Newtonian theory which was otherwise mathematically coherent and compatible with all the relevant empirical data known as the time. However, the need to remove this flaw at the conceptual level led ( at least in hindsight) to the birth of general relativity which in turn predicted new testable results not envisaged by the Newtonian theory. What this story tells us is that a *conceptual dilemma* can become *creative* if it is formulated in a sufficiently precise manner, both mathematically and conceptually. It is this motivation that underpins the present phase of investigations on the foundations of quantum mechanics.

The purpose of this article is to provide some flavour of the significant new developments in this area. Since the scope of this article is rather limited, we will restrict ourselves to the basic ideas and pertinent conceptual aspects, leaving out the mathematical and experimental details for which appropriate references will be given. The central issues at the core of the foundations of quantum mechanics are the quantum measurement problem and quantum nonlocality which will be discussed in Sections 2 and 3 respectively.

## 2 The Quantum Measurement Problem

The question of whether there is a need to go *beyond* the standard interpretation so that *alternative* approaches require to be pursued may seem to be largely dependent on subjective predilections. This is, however, not true. There are compelling physical reasons for suspecting that the standard framework of quantum mechanics is fundamentally inadequate, though its empirical success to date is unquestionably impressive. An outstanding puzzle which underscores a subtle inner inconsistency within the standard framework is the *quantum measurement paradox*. This requires for

---

\*Email: dhom@boseinst.ernet.in.



its satisfactory resolution either a realist interpretation of quantum mechanics or an appropriate modification of the standard formalism or both.

The central problem is that the very occurrence of a definite individual outcome (undeniably a fact of experience) cannot be ensured in an entirely consistent way within the standard framework of quantum mechanics. Accepting the statistical feature of quantum mechanics (that the results of measurement can only be predicted statistically) as inescapable, a fundamental significance is ascribed to probability distributions which are reproducible and quantum mechanically predicted. Then it becomes imperative to guarantee the objective reality of individual outcomes whose collection enables us to test the computable probabilities.

In order to explain the essence of the quantum measurement problem, we first consider the quantum mechanical treatment of a typical measurement process. We turn to a particular example of the formation of an  $\alpha$ -particle track in a set of photographic plates. We briefly indicate the main ingredients, following the presentation given by Bell [2]. We assume a highly simplified model of a photographic plate made up of monoatomic layers of atoms (whose thermal excitations are ignored), each with only one possible excited state. We neglect the possibility of elastic scattering (i.e., the possibility of scattering without excitation) of the incident  $\alpha$ -particle from these atoms.

Let the observed  $\alpha$ -particle originate with momentum  $k_0$  from a source at position  $\vec{r}_0$  and its initial state represented by :

$$\psi_0(\vec{r}) = \frac{\exp(ik_0|\vec{r}-\vec{r}_0|)}{|\vec{r}-\vec{r}_0|} \quad (2.1)$$

If  $\phi_0$  is the ground state of the stack of photographic plates, then the initial combined state is simply  $\Psi_i = \psi_0\phi_0$ . We enumerate the atoms of the stack by  $n(=1,2,3,\dots)$ . Due to interaction with the incident  $\alpha$ -particle, these atoms are excited and subsequently ionized. Let  $\phi(n_1, n_2, n_3, \dots)$  denote a state of the stack where atoms  $n_1, n_2, n_3, \dots$  are excited. After the  $\alpha$ -particle-stack interaction, using the usual multiple-scattering approximation to describe the scattered waves, the final combined state can be written in the form [2]:

$$\begin{aligned} \Psi_f = \sum_N \sum_{n_1, n_2, \dots, n_N} \phi(n_1, n_2, n_3, \dots, n_N) f_N(\mu_N) \times \\ \left[ \frac{\exp(ik_N|\vec{r}-\vec{r}_N|)}{|\vec{r}-\vec{r}_N|} \right] \times f_{N-1}(\mu_{N-1}) \times \left[ \frac{\exp(ik_{N-1}|\vec{r}_N-\vec{r}_{N-1}|)}{(|\vec{r}_N-\vec{r}_{N-1}|)} \right] \times \dots \\ \dots \times \left[ \frac{\exp(ik_0|\vec{r}_1-\vec{r}_0|)}{|\vec{r}_1-\vec{r}_0|} \right] \end{aligned} \quad (2.2)$$

which is a sum over all possible sequences of excitations of  $N$  atoms, with  $\vec{r}_1$  denoting the position of atom  $n_1$ ,  $\vec{r}_2$  of atom  $n_2$ , and so on;  $\mu_n$  is the angle between  $\vec{r}_n - \vec{r}_{n-1}$  and  $\vec{r}_{n+1} - \vec{r}_n$  ( $\vec{r} - \vec{r}_N$  for  $n = N$ );  $f_n(\mu)$  is the inelastic scattering amplitude for an  $\alpha$ -particle of momentum  $k_{n-1}$  incident on a single atom;  $k_n = (k_{n-1}^2 - \varepsilon)^{1/2}$ , where  $\varepsilon$  is a measure of atomic wave functions by using, say, the Born approximation. The form of  $f_n(\mu)$  determines that the atoms ionised by excitation lie approximately on a straight line pointing toward the source of  $\alpha$ -particle. Thus the *correlation* between atomic excitations described by a wave function of the nonfactorisable form (2.2) carries information about the passage of a triggering  $\alpha$ -particle.

A generic feature of all the examples of measurement analysed quantum mechanically, such as the one outlined above, is the following. If a system is initially in a state  $\psi(\psi = a\psi_1 + b\psi_2)$  which is a superposition of two states  $\psi_1$  and  $\psi_2$  that are eigenstates of a dynamical variable which is measured, a general characteristic of its interaction with a measuring device is that it results in a final state of the form

$$\Psi = a\psi_1\Phi_1 + b\psi_2\Phi_2 \quad (2.3)$$

where  $\Phi_1$  and  $\Phi_2$  are mutually orthogonal and macroscopically distinguishable states of the device. It is an ineluctable feature of linear unitary quantum mechanical treatment of *any* measurement process that the final state of system coupled to measuring apparatus has the *entangled* (nonfactorisable) form given by Eq.(2.3).

Origin of the much debated measurement problem [3-9] lies in the measuring of a *pure state* wave function in quantum mechanics which gives rise to an inherent *incompatibility* between a wave function of the form (2.3) and the occurrence of a definite measurement result. A pure state in quantum mechanics means that each member (in this case, a system coupled to an apparatus) of an ensemble described by a pure state  $\Psi$  as given by Eq.(2.3) has the *same* wave function  $\Psi$ . Thus

a pure state in quantum mechanics corresponds to a *homogeneous* ensemble whose members are *indistinguishable*. On the other hand, all measurements culminate in the final ensemble of systems coupled to apparatus which is essentially *heterogeneous*. A heterogeneous ensemble is, however, represented by a mixed state in quantum mechanics. Since within standard quantum mechanics under no unitary time evolution a pure state can evolve into a mixed state (see, for instance, [8] pp. 87-88), *how* to coherently accommodate within quantum mechanics the occurrence of distinguishable outcomes is thus an intriguing “paradox”. Not surprisingly, Weinberg [10] has called this “the most important puzzle in the interpretation of quantum mechanics”.

Though reproducible statistical frequencies of events are quantum mechanically computable, the underpinning concept of an individual definite event is difficult to accommodate within the framework of standard quantum mechanics. It is thus a logical *non sequitor* to speak of probabilities of various outcomes when the very occurrence of an individual outcome is not ensured. Recall that because of the entangled nature of the system-apparatus combined state, it is not permitted in quantum mechanics to assign a separate definite state to any individual apparatus; nevertheless, in any given experimental run, we observe what happens to an individual apparatus coupled with a system subjected to measurement.

The quantum measurement riddle is sometimes dismissed as fundamentally the same as the situation in classical statistical mechanics. This point is argued for example by referring to “a very close analogy between this quantum mechanical problem and the purely classical problem of the tossing of a coin”. Since unknown and uncontrollable elements are inherent in the exact specification of relevant initial conditions, classically the probability for heads is taken as 50%. *Until* we inspect the outcome after tossing a coin, the classical probability is 50% for either outcome. Now if, say, heads is observed, the probability for heads becomes 100%; there is however nothing fundamentally problematic about this feature. As Bell [11] remarks about such classical examples

at least one can envisage an accurate theory to which the restricted account is an approximation. This is *not* possible in quantum mechanics ....It could also be said that even in classical mechanics the human observer is *implicit* for what is interesting if not experienced? But even a human observer is no trouble (in principle) in classical theory - he can be included in the system (in a schematic way) by postulating a “psycho-physical parallelism” - i.e., supposing his experience to be correlated with some functions of the coordinates. This is *not* possible in quantum mechanics, where some kind of observer is not only essential, but *essentially* outside.(italics ours)

This point brings us to one of the crucial components of the measurement problem : An outcome has an objective reality (recorded in terms of changes in the particle properties of the macroscopic measuring device) in the sense of being both intersubjective and out there that can be inspected at will at any instant without perturbing the outcome. Unlike classical mechanics, there is no counterpart of this feature within the theoretical framework of standard quantum mechanics. In other words at the end of a measurement process described quantum mechanically, it is not possible to attribute to a measuring macroapparatus a quantum mechanical representation of its state that embraces a description of all the properties we require it to have. We also mention that the point of the measurement paradox and all the preceding arguments remain entirely unaffected even if the initial state of the macroscopic apparatus is taken to be a mixture of different states rather than a pure state (see Wigner [12] and Leggett [13] for an explicit discussion). For the simplicity of notation we confine ourselves to assigning a pure initial state to the macroapparatus.

The acuteness of this paradox makes us suspect that quantum theory is intrinsically inexact or at least ambiguous at a fundamental level. However, various versions of what may be called the standard approach have tried to alleviate this doubt or minimize this paradox with different arguments.

## 2.1 Variants of the Standard Solution and Their Inadequacies

### 2.1.1 Bohr-Heisenberg Viewpoints

Bohr recognized that if the measuring device were described quantum mechanically, its interaction with the measured system would merely extend the chain of inference without leading to a definite

result. He sought to avoid this problem of infinite regress by decreeing that the interaction between an object and an apparatus is a single unanalyzable whole and the apparatus must be described in classical terms.

An immediate criticism of the preceding point of view is that Bohr makes the very concept of classical measuring instruments “which serve to specify the conditions under which the phenomena appear” [14] depend on an ill-defined limiting procedure. In expressing what Bohr calls “a simple logical demand”, he takes for granted that the composite system in a measurement process can be envisaged as composed of two distinct parts: the measured system and the measuring apparatus. But there is no precise criterion definable within standard quantum mechanics that delineates the borderline between the two. We merely find it convenient to consider some parts of the global system as parts of the instrument or is it determined *a priori* in a more physical way? Is it at all possible within the framework of quantum mechanics to assign a classical description to the measuring apparatus? Such questions remain unanswered in Bohr’s writings.

We now turn to Heisenberg’s advocacy of the Bohrian position. Heisenberg [15] admits that “it is not possible to decide, other than arbitrarily, what objects are to be considered as part of the observed system and what as part of the observer’s apparatus,” but he stresses that once the macroscopic level is reached, since a cut is necessary to avoid the problem of infinite regress of not ensuring a definite outcome, it is of no practical importance where we put the split. According to Bell [16], the Heisenberg dictum was to “put sufficiently much into the quantum system so that the inclusion of more would not significantly alter practical predictions”. Though this recipe is useful, it is ambiguous in principle. There is no fundamental reason why the physics involved in measurements should differ from how other physical interactions are described. Hence the very legitimacy of such a conceptual discontinuity, not so much whether the precise location of its position matters, is the crux of the issue.

In his later years Heisenberg’s position shifted to a subjective approach to the measurement problem. As a corollary to the Bohrian thesis that a measuring device must necessarily be described in terms of classical concepts, Heisenberg inferred that if a measurement interaction is described quantum mechanically it contains “new elements of uncertainty” because “connection with the external world is one of the necessary conditions for the measuring apparatus to perform its function” [17]. In the words of Heisenberg [18], “It is the discontinuous change of our knowledge in the instant of registration that has its image in the discontinuous change of the wave function”.

We highlight certain criticisms. First, the very idea of a change in knowledge obtained through a measurement (not whether such a change corresponds to an actual physical process described by the theory) assumes the occurrence of a definite outcome (otherwise how do we extract relevant measurement information). However a definite outcome is in itself incompatible with the system apparatus joint pure state at the end of a measurement process. It is thus difficult to maintain the logical tenability of the preceding view unless this central problem is considered.

Heisenberg [17] realizes this and notes that to link the mathematical representation of quantum theory “to the question of how real or possible experiments will result” in a definite outcome, we must describe the system apparatus combination in a mixed state. To ensure this, the idea that a macroapparatus is inevitably coupled rather strongly with its environment was introduced [17]. Heisenberg claims that “the compound system of system and measuring apparatus is therefore now described mathematically by a mixture”. The underlying justification rests on contention that since the interaction between an apparatus and its environment (having a large number of degrees of freedom) contains uncontrollable terms that differ from sample to sample, due to incomplete knowledge about the environment, the final state of system and apparatus can be effectively represented by a mixture of states.

Nevertheless the preceding line of reasoning ignores the key fact that whatever the uncontrollable uncertainties in the total Hamiltonian, the entire combination of system and apparatus coupled to environment is bound to be finally left in a nonfactorisable pure state (corresponding to a homogeneous ensemble of systems described by the same composite wave function) under unitary quantum mechanical evolution. See Section 2.1.3 of this article for a proof that irrespective of the details of any interaction, a pure state can evolve unitarily *only* into a pure state.

Hence the crux of the quantum measurement problem has remained unaddressed. One may

of course contend that the reduced density matrix of system and apparatus (obtained by tracing over the environment states) is diagonal. With this assumption Heisenberg [17] suggests using “a statistical mixture in the mathematical representation of the larger system composed of the system and the measuring apparatus.” This is, however, *irrelevant* to the question at issue because the *total* density matrix of system and apparatus and environment is actually *nondiagonal*. It does not matter how small the nondiagonal matrix elements are, so long as they are *not* zero.

### 2.1.2 Decoherence Approach

A popular line of argument (called the decoherence approach) accepts the wave function description (2.3) of the combined state of system and apparatus as a correct and complete representation of the final state but tries to reconcile the occurrence of a definite outcome with the formalism by contending that it is practically impossible (i.e. possible in principle only through extremely complex measurements) to distinguish the pure state denoted by Eq.(2.3) from a statistical mixture of states  $\psi_1\phi_1$  and  $\psi_2\phi_2$ .

Rationale for this argument is examined as follows (see, for example, Zurek [19] and Gottfried [20]); a particularly incisive exposition with an appropriate critique is given by Leggett [3,21]. Due to orthogonality of the apparatus states  $\phi_1$  and  $\phi_2$ , it is clearly not possible to distinguish the pure state (2.3) from the corresponding mixture by measuring the expectation value of a dynamical variable pertaining to the system alone: We must measure an appropriate dynamical variable (say,  $\Omega$ ) of the apparatus as well. Then a necessary condition for discriminating between the pure state (2.3) and the corresponding mixture is that off-diagonal matrix elements  $\Omega_{12}$  and  $\Omega_{21}$  be nonvanishing where:

$$\Omega_{12} = \langle \phi_1 | \Omega | \phi_2 \rangle \quad \Omega_{21} = \langle \phi_2 | \Omega | \phi_1 \rangle \quad (2.4)$$

Bearing in mind that  $\phi_1, \phi_2$  are states of a macroobject comprising a large number of particles, it follows that these matrix elements are nonzero only if the operator  $\Omega$  simultaneously changes the state of a very large number of particles. Typical operators, such as the total momentum or the angular momentum, are sums of single particle operators (not, say, products), and therefore these change the state of only one particle at a time. Thus there is no simple dynamic property of a macroapparatus whose measurement enables us to distinguish the pure state in question from a mixture.

Moreover, the inevitable strong coupling of a macrosystem with its environment gives rise to further difficulties. The apparatus states  $\phi_1, \phi_2$  become entangled with environment states. Consequently, because of the orthogonality of environment states, *no* measurement of a property of the system and apparatus *alone* produces a result other than that expected from a classical mixture of states  $\psi_1\phi_1$  and  $\psi_2\phi_2$ . The only way of discriminating between such a mixture and a pure state (involving environment states) is to measure *correlations* between a macroapparatus and its environment. This is, however, extremely difficult in practice because of the *dissipative* nature of the interaction with the environment, which is fatal to coherence embodied in a pure state. For a detailed articulation of this point, see Leggett [22].

Since it is difficult to envisage a realistic experiment to discern the effects of phase coherence embodied in the pure state description at the end of a measurement process, it is considered natural to assume the pure state behaves *as if* it were a mixed state representing a heterogeneous ensemble of systems coupled with apparatus, where different outcomes correspond to distinct apparatus states. While this point of view has motivated a great deal of elegant theoretical work underscoring the difficulty in observing the effects of interference among macroscopically distinguishable outcomes (for an overview see for example Omnes [23] and van Kampen [24]), so far as solving the basic conceptual problem is concerned, these efforts are unsuccessful.

In spite of the fact that decoherence effects are important in accounting for the usual absence of quantum interference effects in the macrodomain, they are irrelevant as far as the measurement paradox is concerned. This is essentially because the *interpretative shift* from the notion “a pure state of entangled system + apparatus behaves *as if* it were a mixed state” to “a pure state is *actually* a mixed state” entails a major logical *non sequitor* [25]. The fundamental distinction between a pure state and a mixed state inherent in the formalism of quantum mechanics cannot

be ignored by slipping in such an *ad hoc* interpretative drift. According to Bell [26], we do not have here a resolution of the fundamental problem, “but a *change* of the theory at a strategically well chosen point.” The crucial point, as Bell says, is *how* come an *and* is converted into *or*?

It is difficult to accept that the inability to observe interference between macroscopically distinct outcomes (which in principle exist in any coherent superposition) *by itself* authorizes an individual alternative. That is, *how* a pure state is interpreted (corresponding to a homogeneous ensemble of indistinguishable members) cannot be abruptly changed when we reach the *macrolevel* merely because the relevant evidence in terms of interference is inaccessible. The punch line is as Leggett [25] puts it:

can the meaning of the formalism change radically, just because the evidence has disappeared?

For further critiques of the decoherence approach, see for example Home and Whitaker [27], Bell [26], Leggett [3,21].

The logical situation remains essentially unchanged in what is known as the *consistent histories* interpretation which also relies on the decoherence approach. The only difference here is that we discuss state sequences (histories) rather than a state at a single instant. A key assumption is that measured quantities are correlated with decohering histories and only decohering histories can be assigned probabilities [28-32]. This really means that the absence of interference between different alternatives is taken as a justification for a particular alternative. Proponents of this scheme argue that the *correlation* between a microphysical quantity and an appropriate macroclassical variable together with the specification of mathematical conditions ensuring the interference between different histories to be unobservably small for all times that completely define a measurement situation. It is thus obvious that in such a program, the fundamental question as to *how* a particular history is actually realised remains unaddressed.

### 2.1.3 Dirac-von Neumann Projection Postulate

In contrast to that of Bohr or Heisenberg, the Dirac-von Neumann approach seeks to negotiate the measurement riddle by invoking an additional axiom, popularly known as the *projection postulate*. The necessity of adding such a postulate to the quantum formalism was first addressed by Dirac [33] at the 1927 Solvay Congress. Later he introduced a more explicit statement in his famous book [34], viz., “a measurement always causes the system to jump into an eigenstate of the dynamical variable that is being measured, the eigenvalue this eigenstate belongs to being equal to the result of the measurement.” Note that this form of projection postulate is fundamentally inadequate in as much as it eschews the entangled nature of system and apparatus by restricting attention to an abrupt change in the wave function of the measured system alone. If a measurement process is subjected to the Schroedinger evolution, we cannot speak in terms of an independent separate quantum mechanical state assigned to a measured system. This deficiency is remedied in von Neumann’s formulation [35] by postulating the wave function collapse in terms of a dynamic transition from a pure state (of system and apparatus) into a mixed state.

It is straightforward to show that irrespective of the specifics of any interaction, a pure state evolving according to the Schroedinger equation cannot evolve into a mixed state. A convenient mathematical representation of the distinction between a pure state and a mixed state is given in terms of density matrices. A pure state is characterised by  $\rho^2 = \rho$ , whereas for a mixed state,  $\rho^2 \neq \rho$ . Recall that the Schroedinger time evolution of an initial density matrix  $\rho(t_0)$  obeys the following rule :

$$\rho(t) = U(t, t_0)\rho(t_0)U^\dagger(t, t_0) \quad (2.5)$$

where U is unitary satisfying

$$U^\dagger(t, t_0)U(t, t_0) = U(t, t_0)U^\dagger(t, t_0) = I \quad (2.6a)$$

$$U(t_0, t_0) = I \quad (2.6b)$$

The explicit expression for U is given by :

$$U(t, t_0) = \exp[(-i/\hbar)H(t - t_0)] \quad (2.6c)$$

From Eq.(2.5) it follows that :

$$\rho^2(t) = U(t, t_0)\rho(t_0)U^\dagger(t, t_0)U(t, t_0)\rho(t_0)U^\dagger(t, t_0) \quad (2.7)$$

Using Eq.(2.6a)

$$\rho^2(t) = U(t, t_0)\rho^2(t_0)U^\dagger(t, t_0) \quad (2.8)$$

If  $\rho^2(t_0) = \rho(t_0)$ , then using Eq.(2.5) we obtain from Eq.(2.8)

$$\rho^2(t) = \rho(t) \quad (2.9)$$

Thus a pure state can evolve *only* into a pure state through a linear unitary Schroedinger time development. Invoking the projection postulate therefore implies a measurement-induced evolution *not* governed by the Schroedinger equation. To put it more formally, the projection postulate means a *discontinuous nonunitary* change of  $\rho = |\Psi\rangle\langle\Psi|$  (where  $|\Psi\rangle$  denotes the pure state of system and apparatus generated by the Schroedinger treatment of a measurement, say, of the form given by Eq.(2.3) given by

$$\begin{aligned} \rho \rightarrow \rho' = & |a|^2 |\psi_1\rangle\langle\psi_1| |\phi_1\rangle\langle\phi_1| \\ & + |b|^2 |\psi_2\rangle\langle\psi_2| |\phi_2\rangle\langle\phi_2| \end{aligned} \quad (2.10)$$

where  $\rho'$  denotes a mixed state of  $|\psi_1\rangle\langle\psi_1|$  and  $|\psi_2\rangle\langle\psi_2|$  with weight factors  $|a|^2$  and  $|b|^2$ , respectively. Note that in this scheme, the disappearance of coherence in the postmeasurement density matrix  $\rho'$  is regarded as a consequence of the collapse postulate made at the wave function level.

Note that von Neumann recognised that the projection postulate entailed “a peculiar dual nature of the quantum mechanical procedure”. The hypothesis that at some stage during a measurement the Schroedinger evolution requires to be *abruptly* suspended (the putative collapse of a wave function) and an entirely different physical process takes over whose dynamics is unspecified has obvious fundamental difficulties. Apart from the *ad hoc* way this postulate is grafted onto the theory, we are not told *at what point* the collapse occurs or *how long* it takes.

## 2.2 Nonstandard Approaches

Inadequacies of all versions of the standard solution to the measurement problem underscore the need for exploring schemes beyond the so-called standard viewpoints. Broadly speaking such (non-standard) programs can be divided into two classes: approaches accepting the standard formalism as it stands (but these introduce new elements into the conceptual framework) that do not require the idea of wave function collapse in any form and schemes modifying the standard formalism (in a way consistent with empirically verified predictions of the standard formalism) to provide dynamic descriptions of wave function collapse as an objective physical process.

We now discuss three such major (in terms of attention received in the literature) approaches. The first two (the many-worlds approach and Bohm’s causal interpretation) belong to the first class; in the second class we deal with dynamic models for wave function collapse that incorporate stochastic terms into the Schroedinger equation.

### 2.2.1 Many-Worlds Interpretation

The many-worlds interpretation (henceforth abbreviated MWI) confronts the measurement problem by retaining not only the standard formalism but also as much possible of the standard interpretational framework. There are different presentations of MWI, starting from the original version in the 1957 paper by Everett [36]. Here we follow mainly the exposition by Squires [37,38].

While the standard theory results in the superposition of different outcomes given by the system apparatus combined wave function, such as Eq.(2.3), MWI seeks to eliminate the inconsistency between the universal applicability of quantum mechanics and the definiteness of an individual outcome by postulating a special relationship between the final wave function [such as Eq. (2.3)] and the observer’s state of *awareness*. Each term in the final wave function is assumed to correspond to a definite state of awareness registering a particular outcome. Experiences of all the different outcomes are thus considered to be part of the final wave function. This is often expressed graphically by saying that as a result of any observation, the “world” branches into different worlds, and awareness of each outcome belongs to one world. This is the origin of the name *many worlds*. However, as Squires [37,38] emphasizes, here an interpretation is envisaged incorporating perceptions of different outcomes, where each perception involves a relationship between a state of awareness and a corresponding state of the measured system entangled with the apparatus. Using Everett’s terminology we summarise by saying that any *macroscopically discernible* part of the total wave



function in Hilbert space has a perceptible meaning only in relationship to a frame with reference to the observer's mind.

At this stage there is a caveat. The final wave function, such as Eq.(2.3) can be expanded in *any* basis. Therefore the basic tenet of MWI must tacitly assume some *preferred basis* in some suitable macroscopic limit. This is noted by several authors [39-43], although the problem is not addressed in Everett's original paper [36].

In any of its version, pursuing MWI involves a considerable amount of metaphysical baggage. Moreover, there is no strong *a priori* reason for going beyond physics to seek a solution to the quantum measurement problem. A more attractive approach (at least to the majority of the physicists concerned with the measurement problem) is to obtain a solution from within physics unless and until this is shown to be impossible. The most widely discussed two approaches in this direction are discussed in the following two sections. One of them (the Bohmian model) hinges on assuming that the wave function description of the state of an individual system needs to be "completed" by using the concept of "ontological" (i.e., premeasurement and observation-independent) values associated with position variables. The other one seeks to provide a dynamical description of the process of wave function collapse by accepting the premise that a wave function provides a "complete" specification of the state of an individual system. But this approach is based on modifying the Schrodinger equation appropriately. For mathematical details of these approaches and their applications see, for example, the relevant chapters in Home's book [8].

### 2.2.2 Bohmian Approach

Bohm's causal interpretation *reinterprets* the standard formalism of quantum mechanics [44-47] by introducing ontological position variables (say,  $x$ ) that with wave function  $\psi$  provide complete specification of the state of an individual system. In this scheme the measurement problem is resolved by assuming that *all* measurements of microphysical attributes ultimately lead to observations of the *position* of some macroobject serving as an apparatus; in other words, all instrument outputs in the end are assumed to be readings in position space. This is certainly a reasonable assumption, which at least covers a large class of all standard measurements. Even measurements of such physical variables as mass, wavelength, time of flight, which are not directly associated with Hermitian operators, are ultimately measurements of position; for instance, mass is inferred from position in a mass spectrograph, wavelength is obtained from fringe spacing in an interference experiment, and time of flight is inferred from position measurements at different instants.

A definite outcome in an individual measurement is determined by relevant ontological position variables which have well-defined values at all instants. Interpreted in this way the intrinsic inexactness of quantum theory is eliminated by ensuring correspondence between the occurrence of a definite result and functions of spatial coordinates introduced in the theory at a fundamental level. Now recalling the essence of quantum mechanical treatment of a measurement process, note that the final wave function of the system apparatus combination is given by an entangled form with the spatial separation  $\delta y_\beta$  between centres of apparatus wave packets  $\phi_0$  corresponding to different outcomes. For the purpose of perfect measurement,  $\delta y_\beta$  must attain a value significantly larger than the width of the apparatus wave packets  $\phi_0$ , so that these wave packets can be regarded as nonoverlapping in position space.

Each individual outcome is characterised by observing the value of the apparatus variable  $y$ . However, within the standard interpretation, *unless* actually observed,  $y$  is simply an argument of a mathematical function  $\psi$ , devoid of any ontological significance. On the other hand, the outcome of an individual measurement recorded through the value of  $y$  is something to which autonomous physical reality is ascribed. This is automatically ensured in the Bohmian scheme.

Since the measurement interaction causes the configuration space wave function to split into a set of *nonoverlapping* apparatus wave packets, in any single run the apparatus particles enter one of these channels (the Bohmian interpretation legitimizes such terminology by using the notion of particle trajectories). According to this point of view, the fact that apparatus particles enter one of the possible channels leads to a *definite* measurement result. Thus, *in principle*, the final outcome is *causally* related to or is uniquely determined by the premeasurement system-apparatus

wave functions and their initial positions. This is why the Bohmian scheme is often called the *causal interpretation*.

A critical aspect of the Bohmian position may be briefly mentioned here. The separation of the apparatus wave packets occurs in position space and *not* in, say, momentum space. This is conceptually justified, as mentioned earlier, by assuming that *all* measurements eventually reduce to position determinations (e.g., location of a meter needle, distribution of dots in computer print-out). However, it is still formally legitimate to expand the final system-apparatus-combined wave function in any basis. Hence for this viewpoint to be entirely consistent, we need a formal justification for the emergence of a *preferred basis* in the *macroscopic limit*.

As Bell [2] remarks, it is appropriate to refer to the ontological position coordinates as “exposed variables” and to the wave function as a “hidden variable.” It is indeed “ironic that the traditional terminology is the reverse of this.” This situation is somewhat conceptually analogous to that in classical electrodynamics where the abstract notion of fields has a physical manifestation only by the action on charged particles. Similarly it is only by particle attributes recording measurements results that we obtain information about a wave function.

### 2.2.3 Dynamical Models of Spontaneous Wave Function Collapse

The two preceding approaches accept the standard formalism and introduce new ingredients only at the interpretational level. On the other hand if we address the measurement problem by tinkering with the formalism, it is evident from earlier discussions that the modified time evolution must be nonunitary and nonlinear, since a pure state is doomed to remain a pure state under any linear unitary evolution. As already pointed out, the Dirac-von Neumann idea of measurement-induced wave function collapse lacks precision because it assigns a special role to measurement interactions without specifying at what point of complexity an interaction establishing a correlation between the observables of micro and macrosystems becomes a measurement.

Though the transition from micro- to macrosystems is gradual in the actual physical world, differences between the Schroedinger evolution and the Dirac-von Neumann postulated collapse dynamics are rather sharp. Hence it is difficult to comprehend how at some point the linear unitary Schroedinger evolution are suddenly suspended, allowing the collapse dynamics to take over. It is thus clear that a logically coherent scheme for accommodating the notion of wave function collapse must have a seamless mathematical description with no dichotomy between measurement and other interactions; this also means no arbitrary split between micro- and macrosystems. Hence it follows that the collapse process must be *spontaneous* in the sense of being present in the fundamental equation *per se*, without being induced by an external stimulus, such as the system apparatus interaction. In recent years there has been a systematic development of such spontaneous dynamical collapse (henceforth abbreviately denoted by SDC) models. Without going into specifics of these different versions of SDC [48-51] we discuss their general attributes.

Additional nonlinear terms are incorporated into the Schroedinger equation, which entail a modified time evolution of a system. Such new terms in the Schroedinger equation are postulated to satisfy the two diverging desiderata of having a practically negligible effect for all microsystems (a necessary requirement due to the extremely high degree of validity of all tested predictions of the standard quantum formalism in the microdomain) and of being able to induce an appropriately rapid suppression of superpositions of macroscopically distinguishable states in the macrodomain.

Note that this latter feature is required to eliminate quantum mechanical manifestations of superpositions of macroscopically distinct states and also to ensure the definiteness of an individual outcome. Another aspect of these schemes is that the postulated nonlinearity implies a preferred basis of states in ordinary position space, thereby destroying the inherent equivalence between all unitarily related different basis of states in Hilbert space. By incorporating these features various versions of SDC ensure in the following way that a measurement has a definite outcome.

A measurement interaction leads to an entangled state entailing a superposition of macroscopically distinct states of a macroapparatus. During this stage the standard linear form of the Schroedinger equation matters; nonlinear terms are ineffective in the case of the time evolution of a microsystem. Subsequently, the SDC process becomes effective for the macroapparatus, re-



sulting in the disentanglement of superposition provided the following criterion is satisfied: For the occurrence of a definite outcome through the SDC process, relevant *macroscopically distinct* states of an apparatus ( comprising a sufficiently large number of particles) must be localized in position space. Such localised states are mutually separated by distances large enough compared to a suitable microscopic length scale, usually taken to be  $\sim 10^{-5} \text{ cm}$ . Then any superposition of such states is intrinsically unstable, reducing rapidly to any one of the superposed states under the action of SDC induced by nonlinear terms in the evolution equation.

#### 2.2.4 SDC Models versus the Bohmian-Type Scheme

The crucial difference between how these two approaches confront the measurement problem is that in Bohm's scheme the ontological position identifies a definite outcome, but in SDC models the actual physical process of reducing a pure state wave function to a mixed state corresponds to actualizing a definite outcome. Apart from requiring a wave function to specify completely the physical state of an individual system, the SDC framework also implies that an objective physical reality is associated with a wave function to account for the definiteness of a measurement outcome. Yet the particle ontological aspect of recording a measurement outcome (in the sense that in a measurement a particular result is registered in the form of a change in property of some localized element composed of what we call particles) is not covered by an SDC model because a wave function alone is not sufficient for this purpose.

We need an additional particle ontological attribute at a fundamental level, which is precisely what Bohm incorporated into his model. This specific difficulty cannot be addressed in an SDC scheme by simply invoking the formal feature that the wave function is sharply peaked in position space after localization, that is, by literally identifying the localized wave packet with a particle. Such a point of view has difficulties because of the possibility of spreading of a wave function.

Another fundamental difficulty inherent in the SDC approach is referred to as the "problem of tails". This problem is absent in Bohm's scheme. It arises from the fact that in all versions of SDC models, a definite outcome emerging from the dynamical localization process is associated with a collapsed Gaussian wave function of a macroobject, which though sharply peaked in position space, does have *nonzero* tails extending into far away regions. Thus a definite outcome in terms of a definite position is ascribed to an individual system even when its wave function has a nonvanishing component (albeit very tiny) on the eigenmanifold associated with a different outcome.

The above feature has been criticized by a number of authors as undermining SDC models' claim to be a fundamental solution to the measurement problem. The central point of such a criticism is that a position being "almost" defined is *not* the same as a position being defined and an object "almost" being in one state or another is not the same as being in one state or another. The response from proponents of the SDC approach amounts to the following.

A definite outcome corresponding to, say, "a pointer pointing at  $x_1$ " is specified to occur if at least a certain very large percentage (say,  $\alpha$ ) of the probability amplitude (but not necessarily all of it) is concentrated in the "pointing at  $x_1$ " sector of the pointer's configuration space. But then fixing the value of  $\alpha$  is *arbitrary*. There is no unique way of ascertaining whether  $\alpha$  should be 99% or 99.9%, or other. Moreover, what form of correspondence with physical reality in the macrodomain of the  $(1 - \alpha)\%$  of measurements (yielding no definite outcome at all or, in other words, the pointer needle having no definite position) signifies is left *uncertain*. No matter how extremely close to unity  $\alpha$  is chosen, this is certainly a conceptually difficult question for SDC models. Nevertheless, we cannot ignore virtues of the SDC approach which is in line with Bell's aspiration [52] that "one line of development towards greater physical precision would be to have the jump (wave function collapse) in the equations" and not just in words so that "it would come about as a dynamical process in dynamically defined conditions."

*How* to empirically discriminate between the Bohmian approach and SDC-type models is an open question. In fact, it is presently one of the most challenging areas of study to try to conceive ingenious ideas suggesting experimentally realisable situations where such models will give different predictions.

### 3 Quantum Nonlocality

The term nonlocality appears in various guises in the literature. Broadly speaking, it denotes a form of quantum mechanical action at a distance where a distant influence is counterintuitive because of the absence of a classically describable form of physical mediation. At the outset, we must emphasize that any discussion of nonlocality does not take off the ground unless we clearly define the notion of *spatial separation* between relevant physical events involving localized entities. Yet this defies a precise definition if we remain confined within the standard framework of quantum mechanics, where wave functions are *only* abstract symbols for computing observable probabilities.

All sets of basis states in Hilbert space and all Hermitian operators have equal status. Therefore the very formulation of the problem of quantum nonlocality hinges on ascribing a special role to position in ordinary three-dimensional space. The notion of spatial separation in ordinary space has of course a fundamental physical significance related to the fact that all known interactions in nature decrease rapidly with increasing distance. In discussions of nonlocality it is legitimate to take any two localized entities to be sufficiently distant whenever their spatial separation is large enough so that they can be regarded *noninteracting*. Furthermore, a necessary ingredient in this entire issue is the idea of a localized measurement; viz., any event characterising a measurement at a certain instant occurs at a definite position where a macroscopic detector is localized. All manifestations of nonlocality are ultimately discernible through such localized measurements.

In its most general form a nonlocal effect is thus specified in terms of the state of an individual entity being affected by measurements or by any dynamical intervention in a far away localized region of space and time (sufficiently distant but not necessarily a spacelike separation) such that no known physical interaction or influence (propagating in ordinary space and time) can causally connect occurrences in that space and time region to the system in question. As Bell [53] says :

It is the requirement of locality, or more precisely that the result of a measurement on one system be unaffected by operations on a *distant* system with which it has interacted in the *past*, that creates the essential difficulty. (our italics)

Quantum nonlocality differs from nonlocal action entailed by classical nonrelativistic theories (such as Newtonian gravity, electrostatics, heat diffusion) in that it implies action at a distance that does *not* diminish in strength with increasing distance. Basically quantum nonlocality is *kinematic* in nature and pertains to cases where correlation properties embodied in the nonfactorisable quantum mechanical wave functions (the kinematic component of the theory) are *not* fully *reproducible* by a realist theory satisfying the locality condition. This *incompatibility* is *experimentally verifiable* and it can be theoretically demonstrated by using arguments that do not depend on the way a measurement process is described.

#### 3.1 Bell's Theorem

In his famous paper of 1964, Bell [53] formulated a mathematical demonstration of the incompatibility between quantum theory and a broad class of realist or hidden variable theories satisfying the locality condition. Proofs of Bell's theorem are comprehensively reviewed in various places [54-56]. Here we consider a simple version of the proof in the deterministic case. An advantage of this proof over Bell's original proof is that no explicit use is made of the mathematical machinery of hidden variable theories (thereby avoiding an assumption about the distribution function of hidden variables).

A typical example of such proofs pertain to two spin-1/2 particles (1 and 2) in a singlet state, flying apart from each other so that spin is conserved. (This example was first suggested by Bohm [57] for illustrating quantum nonlocality.) Measurements of the components of spins  $\vec{S}_1$  and  $\vec{S}_2$  along different directions are performed on these particles. For any given particle 1, we measure either quantity  $A$  or  $A'$ , where  $A = 2\vec{a} \cdot \vec{S}_1$  and  $A' = 2\vec{a}' \cdot \vec{S}_1$ . (The  $\vec{a}$ ,  $\vec{a}'$  are unit vectors in different directions.) Measured values of  $A$ ,  $A'$  are  $\pm 1$ . Similarly for any given particle 2, the quantity  $B$  or  $B'$  is measured, where  $B = 2\vec{b} \cdot \vec{S}_2$  and  $B' = 2\vec{b}' \cdot \vec{S}_2$  ( $B, B' = \pm 1$ ).

Now consider the combination  $(AB + A'B + AB' - A'B')$ . For any given pair of 1 and 2, we can measure only one of the product quantities  $AB$ ,  $AB'$ ,  $A'B$ ,  $A'B'$ . In each case by construction,

the answer is +1 or -1. The experiment consists in measuring a large number of pairs, with the setting on one side (Particle 1) altered between  $\vec{a}$  and  $\vec{a}'$  and that on the other side (Particle 2) between  $\vec{b}$  and  $\vec{b}'$ . So we have a large number of measurements of each of the quantities  $AB$ ,  $AB'$ ,  $AB'$ , and  $A'B'$ . Basic experimental data are average values of these quantities, which we denote by  $\langle AB \rangle$ , and so on.

We now make the following apparently innocuous set of assumptions :

1. Each *individual* outcome of a measurement is *causally* determined by *supplementary variables* (the so-called hidden variables) that *together* with  $\psi$  *completely* specify the state of an individual quantum entity (*deterministic realist* ingredient).

Note that the notion of hidden variables in the formulation of Bell's theorem is quite general. Hidden variables are regarded as hypothetical parameters determining the outcomes on an *event-by-event* basis.

2. The value of a hidden variable of any given entity after preparing the initial state is *not* affected by events in distant regions of space and time (locality condition).

The conjunction of Assumptions 1 and 2 imply that for each particle there is a definite result predetermined for any observable that is obtained if we measure the observable (This is sometimes referred to as counterfactual definiteness). This result is *not* influenced in any way by measurements in the distant regions of space and time. Let us now see how these suppositions lead to a *testable* constraint on correlation functions. (No input from quantum mechanics is used in the ensuing argument.)

Considering the example of two spin-1/2 particles (1 and 2) in a singlet state, it follows from assumptions 1 and 2 that the hidden variable associated with each particle fixes definite values of both  $A$  (+1 or -1) and  $A'$  (+1 or -1) that are independent of whether  $B$  or  $B'$  is measured on particle 2. Similarly for each particle 2, predetermined definite values of  $B$  and  $B'$  are independent of whether  $A$  or  $A'$  is measured. Consequently each particle pair has a value of either +1 or -1 for each of the quantities  $AB$ ,  $A'B$ ,  $AB'$ ,  $A'B'$ . For each of the 16 different cases corresponding to possible choices of  $\pm 1$  for each  $A$ ,  $A'$ ,  $B$ ,  $B'$  separately :

$$AB + AB' + A'B - A'B' = \pm 2 \quad (3.1)$$

Note that Eq.(3.1) refers to a single pair (or a hypothetical group of pairs corresponding to the same hidden variable specifying their common initial state). The validity of Eq.(3.1) is ensured because both the occurrences of, say,  $A$  in Eq.(3.1) have the *same* value; similar is the case for  $A'$ ,  $B$ , and  $B'$  (locality condition is invoked here).

Summing Eq.(3.2) over the entire ensemble of pairs and taking the average, we obtain

$$|\langle AB \rangle + \langle AB' \rangle + \langle A'B \rangle - \langle A'B' \rangle| \leq 2 \quad (3.2)$$

By virtue of the principle of induction (that the randomly chosen samples of pairs on which the quantities such as  $AB$ ,  $AB'$ , ..... are actually measured are *typical* of the entire ensemble) we can then identify the averages  $\langle AB \rangle$ ,  $\langle AB' \rangle$ , ..... with the experimentally measured values of these quantities. Thus we have a clear-cut prediction for the actual measured quantities given by Eq.(3.2) which is a form of Bell's inequality.

The inequality (3.2) is violated by quantum mechanical results for, say, the singlet state where  $\langle AB \rangle = -\vec{a} \cdot \vec{b} = -\cos\theta$ . The maximum violation occurs when all directions are coplanar, with the  $(\vec{a}, \vec{b})$  angle equal to  $135^\circ$  and the three others to  $45^\circ$ . Then the left hand side of Bell's inequality (3.2) is  $2 \cdot (2)^{1/2}$ . It is more instructive to take  $\vec{a} = \vec{b}$  with  $\vec{a}'$  and  $\vec{b}'$  at an angle  $\theta$  on each side, so that

$$\vec{a} \cdot \vec{b}' = \vec{a}' \cdot \vec{b} = \cos\theta, \quad \vec{a}' \cdot \vec{b}' = \cos 2\theta \quad (3.3)$$

Then the left hand side of Bell's inequality (3.2) becomes  $|-2 - 2\cos\theta(1 - \cos\theta)|$ , which is greater than 2 for any  $\theta < 90^\circ$ .

A remarkable feature of Bell's theorem is that irrespective of the specifics of realist models, it demonstrates that no realist theory satisfying the locality condition can be fully consistent with the quantum mechanical formalism - a classic no-go theorem making it an *empirically decidable* issue whether local realism is tenable in relation to quantum mechanics. This is thus a unique instance where by appealing to experiments on statistical properties of ensembles, we can draw conclusions about assumptions for the putative realist description of individual events.

The experimental refutation of Bell's inequality has far reaching implications even for those who refuse to speculate about realist theories. This is because the validity of quantum correlation functions violating Bell's inequality for *macroscopic separations* between particles imply an *irreducible quantum nonseparability* in its most acute form. By this we mean the repudiation of the cherished notion of *divisibility by thought* which underlines all of classical physics; viz., any composite physical system can be viewed as composed of elements localized in different regions of space, so that knowledge of the full Hamiltonian function and values of relevant basic physical quantities associated with each constituent should yield ipso facto an exhaustive knowledge of the composite system as a whole.

Before proceeding further we discuss an example [58] of correlation in classical physics for which Bell's inequality is always satisfied. The example pertains to a classical object, initially at rest, disintegrating into two spatially separated fragments 1 and 2 carrying angular momenta  $\vec{J}_1$  and  $\vec{J}_2 = -\vec{J}_1$ . The direction of  $\vec{J}_1$  varies randomly over an ensemble of such identical objects. For each fragment we consider measurement of the projection sign of its angular momentum along a certain direction, say,  $r_a = \text{sign}(\vec{J}_1 \cdot \vec{a})$  for 1 and  $r_b = \text{sign}(\vec{J}_2 \cdot \vec{b})$  for 2, where  $\vec{a}$  and  $\vec{b}$  are unit vectors. Possible values for  $r_a$  and  $r_b$  are  $\pm 1$ .

Obviously if  $\vec{a} = \vec{b}$ , we always have  $r_a = -r_b$ , so that  $\langle r_a r_b \rangle = -1$ . Assuming that the direction of  $\vec{J}_1$  is uniformly distributed, the probability that  $r_a r_b = +1$  is  $\theta/\pi$ , and the probability that  $r_a r_b = -1$  is  $(1 - \theta/\pi)$ , where  $\theta$  is the angle between  $\vec{a}$  and  $\vec{b}$  ( $0 \leq \theta \leq \pi$ ). It therefore follows that  $\langle r_a r_b \rangle = (2\theta/\pi) - 1$ , which always satisfies Bell's inequality (3.2). For instance the left hand side of Eq.(3.2) equals exactly 2 for  $\vec{a} = \vec{b}$ ,  $\vec{a} \cdot \vec{b} = \vec{a} \cdot \vec{a} = 1$ , and  $\vec{a} \cdot \vec{b} = \cos\theta$ , and  $\vec{a} \cdot \vec{b} = \cos 2\theta$  [as mentioned earlier in the corresponding quantum case, the left hand side of Eq.(3.2) is larger than 2 for any  $\theta < 90^\circ$ ], indicating that in this classical example the correlation between the angular momenta of the two fragments is as strong as possible for a system that obeys locality. In general the quantum correlation is stronger than or equal to classical correlation. The form of the angular dependence of the quantum correlation function is crucially responsible for the violation of Bell's inequality.

### 3.1.1 Bell's Theorem Using Stochastic Hidden Variables

The demonstration of Bell's theorem in the preceding section is in terms of deterministic hidden variables. Here we discuss its proof based on a general class of stochastic hidden variable models satisfying the locality assumption. (For simplicity and because of its direct relevance to experimental studies, we follow the Clauser-Horne formulation [59] of the probabilistic approach to local realism; a more general formulation is given by Selleri [60]. The basic idea of such hidden variable theories is that the complete hidden variable description of the source does not uniquely determine measured values of local observables pertaining to correlated particles but only probabilities that possible values occur. We can thus assume that individual spin component values along different directions evolve in time stochastically (independent of any measurement), with the complete state of the source controlling only probabilities that particular values will be revealed when subsequent measurements are performed.

The following argument depends on a source emitting two correlated entities 1 and 2 in opposite directions, where two analysers can either transmit or absorb them. The dichotomic choice forced in this way on each quantum entity can be used to define corresponding dichotomic observables by prescribing that  $A(a) = \pm 1$  ( $B(b) = \pm 1$ ), depending on the choice of transmission (+1) or absorption (-1), pertaining to 1(2). This scheme is relevant to actual experiments testing Bell's inequality with pairs of correlated optical photons emitted in atomic cascades. For such photons the binary choice is between transmission and absorption in a polarizer.

Let us invoke a variable to represent the complete physical state of an individual pair of correlated quantum entities (1 and 2) within a general probabilistic scheme in which  $p_1(a, \lambda)$  is the probability that an individual Particle 1 in the state  $\lambda$  crosses the analyser with parameter  $a$ , then is subsequently detected;  $p_2(b, \lambda)$  is the similar probability for 2;  $P(a, b, \lambda)$  represents the joint probability that both 1 and 2 cross their respective analysers with parameters  $a$  and  $b$  and both are detected. The locality condition is expressed by the following factorability condition :

$$P(a, b, \lambda) = p_1(a, \lambda)p_2(b, \lambda) \quad (3.4)$$

with the obvious requirement that the hidden variable distribution function  $\rho(\lambda)$  corresponding to the initial joint state of 1 and 2 and the domain of possible values of  $\lambda$  do not depend on parameter choice for the analyzers. Note that quantities  $P(a, b, \lambda)$ ,  $p_1(a, \lambda)$ ,  $p_2(b, \lambda)$  are defined at the individual level. Observable probabilities at the statistical level (quantum mechanically computable quantities) are expressible as weighted averages of individual probabilities :

$$p_1(a) = \int p_1(a, \lambda)\rho(\lambda)d\lambda \quad (3.5)$$

$$p_2(b) = \int p_2(b, \lambda)\rho(\lambda)d\lambda \quad (3.6)$$

$$P(a, b) = \int P(a, b, \lambda)\rho(\lambda)d\lambda \quad (3.7)$$

To deduce testable constraints on  $p_1(a)$ ,  $p_2(b)$ ,  $P(a, b)$  from the locality condition (3.4) using Eqs. (3.5)-(3.7) we use the following algebraic theorem.

Given real numbers  $x_1, x_2, X, y_1, y_2, Y$  such that

$$0 \leq x_1, x_2 \leq X \quad 0 \leq y_1, y_2 \leq Y \quad (3.8)$$

it follows that

$$-XY \leq x_1y_1 - x_1y_2 + x_2y_1 + x_2y_2 - x_2Y - Xy_1 \leq 0 \quad (3.9)$$

The proof of Eq.(3.9) is straightforward : Since the intermediate quantity in Eq.(3.9) is linear in each of the four variables  $x_1, x_2, y_1$  and  $y_2$ , we seek its extreme on the boundary of these variables. Let us take  $x_1 = p_1(a_1, \lambda)$ ,  $x_2 = p_1(a_2, \lambda)$ ,  $y_1 = p_2(b_1, \lambda)$ ,  $y_2 = p_2(b_2, \lambda)$  and assume that ontological probabilities  $p_{1,2}$  at the individual level (not directly observable) lie between 0 and 1, so that  $X = Y = 1$ .

Then Eq.(3.9) reduces to

$$\begin{aligned} -1 \leq & p_1(a_1, \lambda)p_2(b_1, \lambda) - p_1(a_1, \lambda)p_2(b_2, \lambda) + p_1(a_2, \lambda)p_2(b_1, \lambda) \\ & + p_1(a_2, \lambda)p_2(b_2, \lambda) - p_1(a_2, \lambda) - p_2(b_1, \lambda) \leq 0 \end{aligned} \quad (3.10)$$

By invoking the locality condition (3.4) we can write

$$\begin{aligned} -1 \leq & P(a_1, b_1, \lambda) - P(a_1, b_2, \lambda) + P(a_2, b_1, \lambda) \\ & + P(a_2, b_2, \lambda) - p_1(a_2, \lambda) - p_2(b_1, \lambda) \leq 0 \end{aligned} \quad (3.11)$$

Multiplying by the probability density  $\rho(\lambda)$ , integrating over  $\lambda$ , and using Eqs. (3.5)-(3.7), we obtain

$$\begin{aligned} -1 \leq & P(a_1, b_1) - P(a_1, b_2) + P(a_2, b_1) \\ & + P(a_2, b_2) - p_1(a_2) - p_2(b_1) \leq 0 \end{aligned} \quad (3.12)$$

which is a form of Bell's inequality known as inhomogeneous inequality, since it is based on both double and single detection probabilities.

An important ingredient in the preceding proof is the condition that the probabilities pertaining to the distribution of hidden variables are positive, and not larger than 1. We may argue that since such probabilities are not directly observable, it should not be objectionable to invoke negative probabilities for hidden variables. Then of course Bell's inequality is no longer derivable for stochastic hidden variable models. In fact there are explicit examples [61,62] of stochastic hidden variable models that reproduce quantum mechanical violations of Bell's inequality at the expense of allowing negative probabilities.

General arguments [63,64] show that we can always reproduce quantum mechanical results for nonfactorable state vectors of correlated systems by stochastic hidden variable models using negative probabilities. However, we stress that to formulate hidden variable models, it is conceptually illegitimate to use nonphysical negative probabilities even though computed testable probabilities are ensured to be positive and physically meaningful. This is because, within the framework of a hidden variable model, quantities interpreted as probabilities at the hidden (individual) level have the same physical and ontological status as probabilities actually measurable at the ensemble level through frequencies of repeated events.

### 3.1.2 Different Local Realist Inequalities

Bell's inequality is one of many inequalities that can be deduced from local realism.

The first examples of inequalities providing physical restrictions *not* contained in Bell's inequality were given by Roy and Singh [65]. They showed, for example, that local realism implies

$$\sum \sum C_{ij} P(a_j, b_j) \leq 6 \quad (3.13)$$

where

$$C_{ij} = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & -1 & 1 & 0 \\ 1 & 1 & 0 & -1 & -1 \\ 1 & -1 & 0 & 0 & 0 \end{pmatrix}$$

Suppose for example that

$$P(a_j, b_j) = \frac{1}{3}[2 + C_{ij}(1 - C_{ij})] \quad (3.14)$$

so that  $P(a_j, b_j) = 0$  if  $C_{ij} = -1$  and  $P(a_j, b_j) = (2/3)$  otherwise. We see that Eq.(3.13) is violated, since its left-hand side equals  $(20/3)$ .

Any Bell combinations of four of these correlation functions can instead take only the values  $(6/3)$ ,  $(4/3)$ , or  $0$ ; thus none of the corresponding inequalities is violated. The set of 20 values (3.14) of  $P(a_j, b_j)$  violates local realism, although it satisfies all conceivable Bell inequalities.

Garuccio and Selleri [66] could show that given numerical coefficients  $c_{ij}$  to be real but otherwise arbitrary and correlation functions  $P(a_j, b_j)$  with  $i = 1, \dots, n$  and  $j = 1, \dots, m$ , local realism implies

$$-M_0 \leq \sum \sum c_{ij} P(a_j, b_j) \leq M_0 \quad (3.15)$$

with

$$M_0 = \text{Max}_{\xi, \eta} (\sum \sum c_{ij} \xi_i \eta_j) \quad (3.16)$$

where among all possible choices of the sign factors  $\xi_i = \pm 1 (i = 1, \dots, n)$  and  $\eta_j = \pm 1 (j = 1, \dots, m)$  we must take the one corresponding to the maximum value of the quantity within parenthesis in Eq.(3.16). A detailed account of this development with various proofs is given in Garuccio [67].

### 3.1.3 Quantum Violation of Local Realism in the Macroscopic Limit

The study of quantum nonlocality embodied in the many-particle entangled states assumes an additional significance when examining whether such nonlocality persists in the appropriate *macrolimit*. Mermin[68] initiated investigations on this topic, followed by Roy and Singh [69], who show that the quantum mechanically predicted violation of the N-particle local realist inequality persists for large values of N - in fact, the magnitude of such a violation grows exponentially for large N.

Home and Majumdar [70] showed that there exist entangled N-particle states (each particle having arbitrary spin-j) whose quantum mechanical expectation value of a hermitian operator exceeds the classical realist bound; the violation persists for large values of N and j. All such works help to reinforce the view that if a wave function preserves its quantum character in the macroscopic domain, even large quantum numbers do not guarantee the disappearance of nonclassical effects. In the present case the many-particle entanglement in a wave function is responsible for large scale violation of classical realism in the domain where both the relevant parameters, quantum number j and number of particles N are arbitrarily large.

Note that the type of macroscopic quantum effect alluded to in this section which shows irreducible quantum nonlocality in the appropriate macrolimit is *different* from the macroscopic quantum effects stemming from a coherent superposition of macroscopically distinguishable states where each state is an entangled many-body wave function corresponding to the state of a macroscopic object as a whole.

### 3.1.4 Experimental Tests : Present Status and Some Recent Suggestions

When assessing the experimental tests of Bell-type inequalities, we must distinguish between the following two types of inequalities.

\* Weak Inequality : An inequality deduced from local realism alone, which is incompatible with quantum mechanical results in the case of ideal experiments with highly efficient detectors.

\* Strong Inequality : An inequality deduced from local realism with an additional assumption, which is violated by quantum mechanical results in the case of actual experiments using the available low-efficiency detectors.

Considering typical experimental arrangements employed when testing correlations between polarizations on photons, a standard form of Bell's inequality for this purpose is given by Eq.(3.12).



This equation follows from local realism without an additional assumption. However inequalities of the type in Eq.(3.12) are not violated by the quantum mechanical predictions for available photomultiplier detectors in the visible wavelength range, with low efficiencies  $\simeq 20\text{-}30\%$ . For this reason one common procedure (from Clauser and Horne [59]) is to invoke the following supplementary assumption : For every photon in the state specified by a hidden variable  $\lambda$ , the probability of detection with a polarizer placed before a detector is less than or equal to the detection probability with the polarizer removed. (This is often referred to as the “nonenhancement” assumption).

Note that this is a nontrivial postulate at the level of individual photons with a particular  $\lambda$  (for an in-depth analysis of the nontrivial nature of this hypothesis in the context of local hidden variable models, see Selleri [71], Ferrero et al.[72], and Santos [73]), though it is of course true at the observable level of statistical results when averaged over all  $\lambda$ s.

This additional hypothesis means the following inequalities

$$\begin{aligned} p_1(a_1, \lambda) &\leq p_1(\infty, \lambda) & p_1(a_2, \lambda) &\leq p_1(\infty, \lambda) \\ p_2(b_1, \lambda) &\leq p_2(\infty, \lambda) & p_2(b_2, \lambda) &\leq p_2(\infty, \lambda) \end{aligned} \quad (3.17)$$

where the symbol  $\infty$  indicates that the polarizer was removed.

We use Eq.(3.9) with  $X = p_1(\infty, \lambda)$ ,  $Y = p_2(\infty, \lambda)$  to obtain the following version of strong inequality

$$\begin{aligned} -D_0 &\leq P(a_1, b_1) - P(a_1, b_2) + P(a_2, b_1) \\ &\quad + P(a_2, b_2) - P(a_2, \infty) - P(\infty, b_1) \\ &\leq 0 \end{aligned} \quad (3.18)$$

where  $D_0$  is the double-detection probability in the absence of intervening polarizers and  $P(a_2, \infty)$ ,  $P(\infty, b_1)$  are measured probabilities with the second and the first polarizer removed, respectively. At this stage a numerical example helps us appreciate the difference between inequalities (3.12) and (3.18).

We define

$$\Gamma = P(a_1, b_1) - P(a_1, b_2) + P(a_2, b_1) + P(a_2, b_2) \quad (3.19)$$

Then the inequalities (3.12) and (3.18) can, respectively, be written as

$$-1 + p_1(a_2) + p_2(b_1) \leq \Gamma \leq p_1(a_2) + p_2(b_1) \quad (3.20)$$

$$D_0 + P(a_2, \infty) + P(\infty, b_1) \leq \Gamma \leq P(a_2, \infty) + P(\infty, b_1) \quad (3.21)$$

These should be compared with the quantum mechanical results

$$P(a_1, b_1) = \left(\frac{1}{4}\right) [\varepsilon_+^1 \varepsilon_+^2 + \varepsilon_-^1 \varepsilon_-^2 \cos 2(a_1 - b_1)] \eta_1 \eta_2 \quad (3.22)$$

$$\begin{aligned} p_1(a_1) &= \varepsilon_+^1 \frac{\eta_1}{2} & p_2(b_1) &= \varepsilon_+^2 \frac{\eta_2}{2} \\ D_0 &= \eta_1 \eta_2 & P(a_2, \infty) &= \frac{\varepsilon_+^1 (\eta_1 \eta_2)}{2} \\ P(\infty, b_1) &= \frac{\varepsilon_+^2 (\eta_1 \eta_2)}{2} \end{aligned} \quad (3.23)$$

where  $\eta_1$ ,  $\eta_2$  are efficiencies of the two photomultiplier detectors on both sides and  $\varepsilon_{\pm}^{1,2}$  are parameters determining efficiencies of polarizers 1 and 2 on both sides corresponding to binary choices between transmission (+) and absorption (-) of a photon when passing through a polarizer. Recalling that the typical numerical values of experimental parameters are

$$\begin{aligned} \varepsilon_+^1 &= \varepsilon_+^2 \simeq 1 & \varepsilon_-^1 &= \varepsilon_-^2 \simeq 1 \\ \eta_1 &= \eta_2 \simeq 0.2 - 0.3 \end{aligned} \quad (3.24)$$

Inequalities (3.20) and (3.21) reduce, respectively, to

$$-0.8 \leq \Gamma \leq 0.2 \quad (3.25)$$

$$0 \leq \Gamma \leq 0.08 \quad (3.26)$$

Local realism gives the set of possible values of  $\Gamma$  a spread of about 1.0, while the additional assumption reduces this figure to about 0.08. Using Eq.(3.19), quantum mechanical results (3.22), and actual experimental parameters (3.24), we can compute  $(\Gamma)_{min}$ , which leads to  $(\Gamma)_{min} = -0.00138$ . This is consistent with the weak inequality (3.25), but it violates the strong inequality (3.26). It is therefore arguable that this observed disagreement stems essentially from the additional postulate used in deriving the strong inequality, a postulate that could then be interpreted as inapplicable at the level of hidden variables  $\lambda$ s. In fact explicit local hidden variable models (albeit ad hoc) are formulated that reproduce quantum mechanical results (3.22) and (3.23) for the actual experimental parameters (3.24) but at the expense of violating the additional assumption [74, 75].

Since this assumption at the level of unobservable  $\lambda$ s cannot be tested directly, one way of making further progress is to subject its various implications to appropriate scrutiny.

Apart from using atomic cascade sources, a number of tests of Bells' inequality have been performed [76,77] using the parametric down-conversion technique for producing photon pairs correlated in polarization. (A laser beam produces a pair of degenerate down-converted photons in a nonlinear crystal of potassium dihydrogen phosphate.) Again these experiments have the same low-efficiency problem mentioned earlier.

Because of the importance of these conceptual issues we need improved tests of Bell-type inequalities based on various quantum systems. Examples involving the decay of a  $J^{PC} = 1^{--}$  state (for example spin-1  $\Phi$  resonance or spin-1  $\Upsilon(4s)$  vector meson) into a pair of neutral pseudoscalar meson-antimeson  $M^0 - \bar{M}^0$  (for example  $k^0 - \bar{k}^0$ ) have been investigated by various authors [78-83].

#### Bell-Type Inequalities

### 3.2 Demonstration of Quantum Nonlocality without Using Bell-Type Inequalities

Greenberger, Horne, and Zeilinger [84] (henceforth GHZ) initiated a new line of study of quantum nonlocality by demonstrating an incompatibility between quantum mechanics and local realism in a nonstatistical way without using Bell-type inequalities. In its simplest version, the argument develops by using a system of three spin-(1/2) particles, mutually correlated and spatially separated; on each of these measurements of x or y spin components are considered. Let the composite state be represented by (referring to z-axis components)

$$|\Psi\rangle = \left(\frac{1}{2}\right)^{1/2} (|1, 1, 1\rangle - |-1, -1, -1\rangle) \quad (3.27)$$

Here  $|1, 1, 1\rangle$  and  $|-1, -1, -1\rangle$  denote states for which  $\sigma_z$  eigenvalues of the three particles are all  $= +1$  or  $-1$ , respectively. It is easily verifiable by direct calculation that  $|\Psi\rangle$  given by Eq.(3.27) satisfies the following eigenvalue equations

$$\sigma_x^1 \sigma_x^2 \sigma_x^3 |\Psi\rangle = -|\Psi\rangle \quad (3.28a)$$

$$\sigma_x^1 \sigma_y^2 \sigma_y^3 |\Psi\rangle = |\Psi\rangle \quad (3.28b)$$

$$\sigma_y^1 \sigma_x^2 \sigma_y^3 |\Psi\rangle = |\Psi\rangle \quad (3.28c)$$

$$\sigma_y^1 \sigma_y^2 \sigma_x^3 |\Psi\rangle = |\Psi\rangle \quad (3.28d)$$

where superscripts 1-3 designate the particles, respectively. In this example it is possible to determine any  $\sigma_x^1, \sigma_y^1, \sigma_x^2, \dots$  by distant measurements on the other two particles (e.g., to know  $\sigma_x^1$ , we need only measure  $\sigma_y^2$  and  $\sigma_y^3$  at distant locations). Applying the notion of local realism it may thus seem legitimate to assume that observed individual values of  $\sigma$ -operators are predetermined (by, say, hidden variables) and that any such individual value is independent of whichever sets of three single-particle spin measurements we choose to make on these spatially separated particles; we call these values  $m_x^1, m_y^1, m_x^2, \dots$ .

Consistent with Eq.(3.28a-d) we then have the following relations for any set of three such correlated particles whose m values are determined by fixed hidden variables:

$$m_x^1 m_x^2 m_x^3 = -1 \quad (3.29a)$$

$$m_x^1 m_y^2 m_y^3 = +1 \quad (3.29b)$$

$$m_y^1 m_x^2 m_y^3 = +1 \quad (3.29c)$$

$$m_y^1 m_y^2 m_x^3 = +1 \quad (3.29d)$$

Note that in the preceding relations the notion of local realism implies that the individual value of any one of the quantities  $m_{x,y}^{1,2,3}$  is the same irrespective of the equation in which it occurs [e.g., the value of  $m_x^1$  is the same in Eq.(3.29a-d)]. Then Eq.(3.29a-d) is not mutually consistent, since from Eq.(3.29a-d) we obtain (recall that  $m_{x,y}^{1,2,3} = \pm 1$ ):

$$m_x^1 m_x^2 m_x^3 = +1$$

which contradicts Eq.(3.29a). This demonstrates that for multiparticle states of the type in Eq.(3.27), quantum mechanical predictions are incompatible with local realism. The chief merit of this form of argument is its nonstatistical nature: We are not concerned with measurement statistics involving questions about the size of the relevant ensemble, statistical fluctuations, and so on. The GHZ argument reveals that the notion of local realism is inconsistent with quantum



mechanics even in the case of perfect correlations (i.e, even at the maximum correlation angles  $\theta = 0$  and  $\theta = \pi$  for which quantum mechanics makes nonprobabilistic, definitive predictions regarding the correlation properties).

Bell-type arguments demonstrate an incompatibility for imperfect correlations only when the specific form of the correlation as a function of the relative orientation between measurement axes plays a central role. The GHZ argument is also concerned with measured values only along orthogonal axes (say, x or y spin components) whereas we require measurements along various nonorthogonal directions to demonstrate quantum violation of Bell's inequality.

There exist other arguments as well which demonstrate, independent of Bell-type inequalities, that local realism and quantum mechanics are mutually incompatible. See, for example,[85].

### 3.3 Quantum Teleportation

A recent scheme suggested by Bennett et al.[86] illustrates quantum nonlocality in a rather striking way involving an interplay between entanglement of states (kinematics) and measurement-induced effect on such an entanglement. This approach enables the state of a given particle to be prepared *identical* to the state of a distant particle with which no direct interaction occurs. Of late, this curious way of transporting the quantum state of an object without requiring the object itself to be transported is attracting much attention, particularly because of recent claims to have experimentally realized this process [87-89].

Classically, moving an object means moving all the particles it is made of. However, an "object" in quantum mechanics is solely characterized by the quantum state of the particles it is made of. Thus reconstructing the quantum state of a particle out of the state of a distant particle is a form of "transportation" of an "object" in a quantum mechanical sense.

In the method of quantum teleportation, a pair of given quantum systems (designated by, say, 2 and 3) is known to be prepared in an entangled state D and distributed between two spatially separated partners, say, Alice and Bob. Suppose Alice has also a test system (designated by, say, 1) in a particular state that may be unknown to her. This approach transfers the state of System 1 itself. The correlated state D of Systems 2 and 3 provides the *quantum channel* crucial for this process. Of course, we cannot preserve the original state 1 intact - it changes in the process. Also note that this form of teleportation of a quantum state cannot take place instantaneously or across spacelike separation because apart from the quantum correlation, it requires sending a classical message (say, a telephone call) from Alice to Bob.

We explain the method by taking all three systems concerned to be spin-1/2 particles and D to be the singlet state given by

$$|\psi_{23}\rangle = \left(\frac{1}{2}\right)^{1/2} (|\uparrow\rangle_2 |\downarrow\rangle_3 - |\downarrow\rangle_2 |\uparrow\rangle_3) \quad (3.33)$$

where  $|\uparrow\rangle_i$  and  $|\downarrow\rangle_i$  are eigenstates of the spin operator  $\sigma_i$  and subscripts 2 and 3 label particles that are with Alice and Bob, respectively. The state of particle 1 with Alice, which is to be transported, is denoted by  $|\phi\rangle_1$ , written as

$$|\phi\rangle_1 = a |\uparrow\rangle_1 + b |\downarrow\rangle_1 \quad (3.34)$$

where  $|a|^2 + |b|^2 = 1$ .  $|\phi\rangle_1$  is supposed to be unknown.

The initial state of the entire system comprising Particle 1 and the correlated pair 2+3 is given by

$$|\psi_{123}\rangle = |\phi\rangle_1 |\psi_{23}\rangle \quad (3.35)$$

Using Eqs. (3.33) and (3.34), we can rewrite Eq.(3.35) as:

$$|\psi_{123}\rangle = \frac{a}{(2)^{1/2}} (|\uparrow\rangle_1 |\uparrow\rangle_2 |\downarrow\rangle_3 - |\uparrow\rangle_1 |\downarrow\rangle_2 |\uparrow\rangle_3) + \frac{b}{(2)^{1/2}} (|\downarrow\rangle_1 |\uparrow\rangle_2 |\downarrow\rangle_3 - |\downarrow\rangle_1 |\downarrow\rangle_2 |\uparrow\rangle_3) \quad (3.36)$$

Next we write each direct product state of 1 and 2 in terms of the following basis vectors :

$$|\psi_{12}^+\rangle = (1/2)^{1/2} (|\uparrow\rangle_1 |\downarrow\rangle_2 + |\downarrow\rangle_1 |\uparrow\rangle_2) \quad (3.37a)$$

$$|\psi_{12}^-\rangle = (1/2)^{1/2} (|\uparrow\rangle_1 |\downarrow\rangle_2 - |\downarrow\rangle_1 |\uparrow\rangle_2) \quad (3.37b)$$

$$|\phi_{12}^+\rangle = (1/2)^{1/2} (|\uparrow\rangle_1 |\uparrow\rangle_2 + |\downarrow\rangle_1 |\downarrow\rangle_2) \quad (3.37c)$$

$$|\phi_{12}^-\rangle = (1/2)^{1/2} (|\uparrow\rangle_1 |\uparrow\rangle_2 - |\downarrow\rangle_1 |\downarrow\rangle_2) \quad (3.37d)$$

which form a complete orthogonal basis for 1 and 2. Using Eq.(3.37a-d),  $|\psi_{123}\rangle$  given by Eq.(3.36) can be expressed as:

$$\begin{aligned} |\psi_{123}\rangle = & \left(\frac{1}{2}\right) [(-a|\uparrow\rangle_3 + b|\downarrow\rangle_3)|\psi_{12}^+\rangle \\ & + (-a|\uparrow\rangle_3 - b|\downarrow\rangle_3)|\psi_{12}^-\rangle + (a|\downarrow\rangle_3 - b|\uparrow\rangle_3)|\phi_{12}^+\rangle \\ & + (a|\downarrow\rangle_3 + b|\uparrow\rangle_3)|\phi_{12}^-\rangle] \end{aligned} \quad (3.38)$$

If Alice performs measurements on the joint system of 1 and 2 to distinguish between states  $|\psi_{12}^\pm\rangle, |\phi_{12}^\pm\rangle$ , then the final combined state is given by

$$\begin{aligned} |\Psi\rangle = & (1/2)[(-a|\uparrow\rangle_3 + b|\downarrow\rangle_3)|\psi_{12}^+\rangle|A_1\rangle \\ & + (-a|\uparrow\rangle_3 - b|\downarrow\rangle_3)|\psi_{12}^-\rangle|A_2\rangle \\ & + (a|\downarrow\rangle_3 - b|\uparrow\rangle_3)|\phi_{12}^+\rangle|A_3\rangle \\ & + (a|\downarrow\rangle_3 + b|\uparrow\rangle_3)|\phi_{12}^-\rangle|A_4\rangle] \end{aligned} \quad (3.39)$$

where  $|A_1\rangle, |A_2\rangle, |A_3\rangle, |A_4\rangle$  are mutually orthogonal states corresponding to different distinguishable outcomes; Eq.(3.39) is thus an *incoherent* sum of four terms thereby indicating *disentanglement* of the state  $|\psi_{23}\rangle$  given by Eq.(3.33). It therefore follows that for each measurement outcome, Bob's Particle 3 is left in one of the four equally likely states

$$\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}, \quad -\begin{pmatrix} a \\ b \end{pmatrix}, \quad \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} \quad (3.40)$$

Recalling that the state  $|\phi\rangle_1$  in Eq.(3.34) to be transported is given by  $\begin{pmatrix} a \\ b \end{pmatrix}$ , then each of the states in Eq.(3.40) is unitarily related to  $\begin{pmatrix} a \\ b \end{pmatrix}$  by a rotation of  $\pi$  around the x, y, z axes. [In the case corresponding to the apparatus state  $|A_2\rangle$ , no such rotation is required because the state of 3 is the same as Eq.(3.34) except for an irrelevant phase factor.] Thus whatever be the outcome of measurement performed by Alice, teleportation of state to particle 3 with Bob is achieved in each case by transmitting the measurement result to Bob through an ordinary classical channel, such as by a telephone call. Afterward Bob applies the required rotation to transform the state of his particle into a replica of the original state (3.34) of Particle 1. At the end of this process, Alice has no trace of the state in Eq.(3.34); instead particles 1 and 2 are left in one of the joint states  $|\psi_{12}^\pm\rangle$  or  $|\phi_{12}^\pm\rangle$ .

Note that in the argument demonstrating the possibility of quantum teleportation it is not necessary to assume wave function collapse; for convenience, many authors discuss teleportation by explicitly invoking the hypothesis of collapse. Incoherence between the four terms in Eq.(3.39) (due to mutual orthogonality of the apparatus states) suffices to ensure that Particle 3 is effectively left in one of the required states enabling the teleportation to be accomplished. We emphasize that quantum correlations embodied in the joint wave function of 2 and 3, together with appropriately chosen measurements on 1 and 2, result in *disentanglement* of the entangled wave function of 2 and 3 which leads to the predicted teleportation of a quantum state.

## 4 Concluding Remarks

For the sake of brevity we have refrained from discussing here a number of significant areas of research on the foundational issues of quantum mechanics. For example, we have not dwelled on the recent investigations throwing new light on wave - particle duality and Bohr's complementarity principle, theoretical and experimental studies uncovering new types of quantum effect such as the Quantum Zeno Effect which has far reaching conceptual ramifications. For elaborate discussions of these topics see the relevant chapters in Home's book [8].

Studies in recent years have not only sharpened our understanding of the conceptual inadequacies of the standard model of quantum mechanics but, more importantly, a number of alternative approaches have been developed which seek to alleviate deficiencies of the standard interpretation. As discussed in this article, such *nonstandard* schemes are not just qualitative, but are also subject to precise *quantitative* formulations. The thrust area at this stage is to investigate the possibility of their testable predictions beyond the standard framework and whether some of these models

may be empirically discriminated from one another. Of course, conceptual as well as mathematical refinements of these schemes are necessary in order to reach a stage when decisive empirical judgements will be possible.

Studies to this end not only deepen our understanding of the “vulnerable” areas of quantum mechanics but also reveal hitherto unexplored facets of the theory. For instance, the empirically relevant manifestations of quantum entanglement that have been discovered related to quantum nonlocality have of late acquired considerable significance in view of their applications in the context of cryptography, quantum communications and quantum computing [90]. Novel approaches have been developed to gain fresh insights into the quantum measurement problem which will make it empirically investigable, like biomolecular systems being used as mesoscopic quantum measuring devices [91]. Ingenious new models of quantum mechanics have also been proposed which have the potentiality of yielding predictions not contained within the standard framework [92].

To conclude, we may quote Bell [93] : “It seems to me possible that the continuing anxiety about what quantum mechanics means or entails will lead to still more tricky experiments which will eventually find some soft spot”.

## References

- [1] A.Shimony in *The New Physics*, edited by P.C.W. Davies, Cambridge University Press, Cambridge (1989).
- [2] J.S.Bell, in *Speakable and Unspeakable in Quantum Mechanics* (Cambridge University Press, Cambridge, 1987), pp. 120-21.
- [3] A.J.Leggett, in *Quantum Implications* ( B.J. Hiley, & D. Peat, eds.)(Routledge & Kegan Paul, London, 1987), pp. 85-104.
- [4] B.d’Espagnat, *Veiled Reality - An Analysis of Present -Day Quantum Mechanical Concepts* (Addison - Wesley, Reading, 1994).
- [5] J.T.Cushing, *Quantum Mechanics - Historical Contingency and the Copenhagen Hegemony* (University of Chicago Press, Chicago, 1994), chap.3.
- [6] A.Whitaker, *Einstein, Bohr and the Quantum Dilemma* (Cambridge University Press, Cambridge, 1996).
- [7] M.Namiki, S.Pascasio, and H.Nakazato, *Decoherence and Quantum Measurement* (World Scientific, Singapore, 1997).
- [8] D.Home, *Conceptual Foundations of Quantum Physics - An Overview from Modern Perspectives* (Plenum Press, New York, 1997).
- [9] P.Mittelstaedt, *The Interpretation of Quantum Mechanics and the Measurement Process* (Cambridge University Press, Cambridge 1998).
- [10] S.Weinberg, *Dreams of a Final Theory*, (Vintage, London, 1993), p. 64.
- [11] J.S.Bell, in *Speakable and Unspeakable in Quantum Mechanics* (Cambridge University Press, Cambridge, 1987), p. 125.
- [12] E.P.Wigner, *Am. J. Phys.* **31**, 6 (1963).
- [13] A.J.Leggett, in *Lesson of Quantum Theory* (J. de Boer, E. Dal, and O. Ulfbeck, eds.)(Elsevier, Amsterdam, 1986), p. 47.
- [14] N.Bohr, *Dialectica* **2**, 312 (1948).
- [15] W.Heisenberg, *Physical Principles of the Quantum Theory* (University of Chicago Press, Chicago, 1930; reprinted Dover, New York), p. 64.

- [16] J.S.Bell, in *Speakable and Unspeakable in Quantum Mechanics* (Cambridge University Press, Cambridge, 1987), p. 124.
- [17] W.Heisenberg, in *Niels Bohr and the Development of Physics* (W. Pauli, ed.)(Pergamon Press, Oxford, UK, 1955), pp.12-29.
- [18] W.Heisenberg, *Physics and Philosophy* (Harper and Row, New York, 1962), chap.3.
- [19] W.H.Zurek, *Prog. Theor. Phys.* **89**, 281 (1993).
- [20] K.Gottfried, *Phys. World* **4**, 10, 34 (1991).
- [21] A.J.Leggett, *Contemp. Phys.* **25**, 583 (1984).
- [22] A.J.Leggett, in *Proc. 1st Int. Symp. Foundations of Quantum Mechanics in the Light of New Technology* (Physical Society of Japan, Tokyo, 1984); *Prog. Theor. Phys. Suppl.* , no. 69, 80 (1980).
- [23] R.Omnès, *Interpretation of Quantum Mechanics* (Princeton University Press, Princeton, NJ, 1994), chap.7.
- [24] N.G.van Kampen, *Physica A*, **153**, 97 (1988).
- [25] A.J. Leggett, *Curr. Sci.* **67**, 785 (1994).
- [26] D.Home and M.A.B.Whitaker, *Phys. Rep.* **210**, 223 (1992), sects. 5.4 and 6.2.
- [27] J.S.Bell, *Phys. World* **3**, no.8, 33 (1990); also in : *Sixty-Two Years of Uncertainty* (A.J. Miller, ed.) (Plenum, New York, 1990), pp. 17-31.
- [28] D.Home and M.A.B.Whitaker, *Phys. Rep.* **210**, 223 (1992), sects. 5.4 and 6.2.
- [29] R.B.Griffiths, *J. Stat. Phys.* **36**, 219 (1984); in *New Techniques and Ideas in Quantum Measurement Theory* (D.M. Greenberger, ed.) (New York Academy of Sciences, New York, 1986); *Found. Phys.* **23**, 1601 (1993).
- [30] R.Omnès, *J. Stat. Phys.* **53**, 893 (1988); *Rev. Mod. Phys.* **64**, 339 (1992).
- [31] M.Gell-Mann and J.B.Hartle, in *Complexity, Entropy, and the Physics of Information* (W. Zurek, ed.) (Addison-Wesley, Reading, MA, 1990); in *Proc. 3d Int. Symp. Foundations of Quantum Mechanics in the Light of New Technology* (S. Kobayashi et al., eds.)(Phys. Soc. of Japan, Tokyo, 1990); *Phys. Rev. D* **47**, 3345 (1993).
- [32] B.d'Espagnat, *J. Stat. Phys.* **56**, 747 (1989); *Found.Phys.* **20**, 1147 (1990).
- [33] . M.Gell-Mann, *Quark and the Jaguar* (Little Brown and Co., London, 1994), pp. 153-54.
- [34] P.A.M.Dirac, in *Electrons et Photons - Rapports et Discussions du Cinquieme Conseil de Physique tenu a Bruxelles 1927* (Gauthier-Villars, Paris, 1928).
- [35] P.A.M.Dirac, *Principles of Quantum Mechanics* (Clarendon, Oxford, UK, 1930), p.36
- [36] J.von Neumann, *Die Mathematische Grundlagen der Quantenmechanik* (Springer-Verlag, Berlin, 1932); English translation : *Mathematical Foundations of Quantum Mechanics* (Princeton University Press, Princeton, NJ, 1955).
- [37] H.Everett, *Rev. Mod. Phys.* **29**, 454 (1957).
- [38] E.J.Squires, in *Quantum Theory without Reduction* (M.Cini and J.M.Levy-Leblond, eds.) (Adam-Hilger, Bristol, UK, 1989).
- [39] E.J.Squires, *Synthese* **97**, 109 (1993).

- [40] B.S.Dewitt and N.Graham, *Many-Worlds Interpretation of Quantum Mechanics* (Princeton University Press, Princeton, NJ, 1973), pp. 155-65.
- [41] L.E.Ballentine, *Found. Phys.* **3**, 229 (1973).
- [42] D.Deutsch, *Int. J. Theor. Phys.* **24**, 1 (1985).
- [43] M.Lockwood, *Mind, Brain, and the Quantum* (Basil Blackwell, Oxford, UK, 1989), chap. 13.
- [44] D.Albert and B.Loewer, *Synthese* **77**, 195 (1988).
- [45] P.Holland, *The Quantum Theory of Motion* (Cambridge University Press, Cambridge, 1993).
- [46] D.Bohm and B.J.Hiley, *The Undivided Universe* (Routledge, London, 1993).
- [47] J.T.Cushing, *Quantum Mechanics - Historical Contingency and the Copenhagen Hegemony* (University of Chicago Press, Chicago, 1994).
- [48] *Bohmian Mechanics and Quantum Theory : An Appraisal* (J.T.Cushing, A.Fine and S.Goldstein, eds.)(Kluwer Academic Publishers, Dordrecht, 1996).
- [49] G.C.Ghirardi, A.Rimini and T.Weber, in *Quantum Probability and Applications* (L.Accardi, and W.von Waldenfels. eds.) (Springer-Verlag, Berlin, 1985), pp. 223-32.
- [50] G.C.Ghirardi, A.Rimini and T.Weber, *Phys. Rev. D* **34**, 470 (1986).
- [51] E. J.Squires, *Phys. Lett. A* **158**, 431 (1991).
- [52] P.Pearle and E.J.Squires, *Phys. Rev. Lett.* **73**, 1 (1994).
- [53] J.S.Bell, in *Speakable and Unspeakable in Quantum Mechanics* (Cambridge University Press, Cambridge, 1987), p.117.
- [54] J.S.Bell, *Physics* **1**, 195 (1964); reprinted in *Speakable and Unspeakable in Quantum Mechanics* (Cambridge University Press, Cambridge, UK, 1987), p.14.
- [55] J.F.Clauser and A.Shimony, *Rep. Prog. Phys.* **41**, 1881 (1978).
- [56] *Quantum Mechanics Versus Local Realism : The Einstein-Podolsky-Rosen Paradox* (F.Selleri, ed.)(Plenum, New York, 1988).
- [57] D.Home and F.Selleri, *Riv. Nuov. Cim.* **14**, no. 9, 1(1991).
- [58] D.Bohm, *Quantum Theory* (Prentice-Hall, Engle hood Cliffs, NJ, 1951), p.614.
- [59] A.Peres, *Am. J. Phys.* **46**, 745 (1978).
- [60] J.F.Clauser and M.A.Horne, *Phys. Rev. D* **10**, 526 (1974).
- [61] F.Selleri, in *Microphysical Reality and Quantum Formalism* (A.van der Merwe et al. eds.)(Kluwer, Dordrecht, Netherlands, 1988).
- [62] J.D.Ivanovic, *Lett. Nuov. Cim.* **22**, 14 (1978).
- [63] W.Muckenheimer, *Lett. Nuov. Cim.* **35**, 300 (1982).
- [64] D.Home and V.L.Lepore, and F.Selleri, *Phys. Lett. A* **158**, 357 (1991).
- [65] G.S.Agarwal, D.Home and W.Schleich, *Phys. Lett. A* **170**, 359 (1992).
- [66] S.M.Roy and V.Singh, *J. Phys. A* **11**, L167 (1978); *J. Phys. A* **12**, 1003 (1979).
- [67] A.Garuccio and F.Selleri, *Found. Phys.* **10**, 209 (1980).

- [68] A.Garuccio, in *Quantum Mechanics versus Local Realism : The Einstein-Podolsky-Rosen Paradox* (F.Selleri, ed.)(Plenum, New York, 1988).
- [69] N.D.Mermin, *Phys. Rev. Lett.* **65**, 1838 (1990).
- [70] S.M.Roy and V.Singh, *Phys. Rev. Lett.* **67**, 2761 (1991).
- [71] D.Home and A.S.Majumdar, *Phys.Rev.A* **52**,4959(1995).
- [72] F.Selleri, *Quantum Paradoxes and Physical Reality* (Kluwer, Dordrecht, Netherlands, 1990).
- [73] M.Ferrero, T.W.Marshall and E.Santos, *Am. J. Phys.* **58**, 683 (1990).
- [74] E.Santos, *Phys. Lett. A* **139**, 431 (1989); *Found. Phys.* **21**, 221 (1991); *Phys. Lett. A* **212**, 10 (1996).
- [75] E.Santos, T.W.Marshall, and F.Selleri, *Phys. Lett. A* **98**, 5 (1983).
- [76] D.Home and T.W.Marshall, *Phys. Lett. A* **113**, 183 (1985).
- [77] Z.Y.Ou and L.Mandel, *Phys. Rev. Lett.* **61**, 50 (1988); Y.H.Shih and C.O.Alley, *Phys. Rev. Lett.* **61**, 2921 (1988).
- [78] Z.Y.Ou, C.K.Hong and L.Mandel, *Opt. Commun.* **67**, 159 (1988); T.E.Kiess, Y.H.Shih, A.V.Sergienko, and C.O.Alley, *Phys. Rev. Lett.* **71**, 3893 (1993).
- [79] J.Six, *Phys. Lett. B* **114**, 200 (1982).
- [80] F.Selleri, *Nuov. Cim. Lett.* **36**, 521 (1983).
- [81] A.Datta and D.Home, *Phys. Lett. A* **119**, 3 (1986).
- [82] D.Home, in *Proc. 3rd Int. Symp. Foundations of Quantum Mechanics* (S.Kobayashi et al., eds.) (Physical Society of Japan, Tokyo, 1990), pp. 43-50.
- [83] A.Datta and D.Home, *Found. Phys. Lett.* **4**, 165 (1991).
- [84] A.Datta, in *Proc. Workshop on High-Energy Physics Phenomenology IV* (A.Datta et al., eds.) (Allied Publishers , New Delhi, 1997).
- [85] D.M.Greenberger, M.A.Horne, and A.Zeilinger, in *Bell's Theorem, Quantum Theory, and Conceptions of the Universe* (M.Kafatos, ed.) (Kluwer, Dordrecht, Netherlands, 1989), pp. 73-76; *Am. J. Phys.* **58**, 1131 (1990); N.D.Mermin, *Am. J. Phys.* **58**, 731 (1990).
- [86] L.Hardy, *Phys. Rev. Lett.* **68**, 2981(1992); **71**, 1665 (1993).
- [87] C.H.Bennett, G.Brassard, C.Crepeau, R.Jozsa, A.Peres, and W.K. Wootters, *Phys. Rev. Lett.* **70**, 1895 (1993).
- [88] D.Boschi, S.Branca, F.de Martini, L.Hardy and S.Popescu, *Phys. Rev. Lett.* **80**, 1121 (1997).
- [89] D.Bouwmeester, J.W.Pan, K.Mattle, M.Eibl, H.Weinfurter, A.Zeilinger, *Nature* **390**, 575 (1997); J.W.Pan, D.Bouwmeester, H.Wienfurter, and A.Zeilinger *Phys. Rev. Lett.* **80**, 3891 (1998).
- [90] S.L.Braunstein, H.J.Kimble, *Nature* **394**, 840 (1998); D.Bouwmeester, J.W.Pan, K.Mattle, M.Eibl, H.Weinfurter, A.Zeilinger, *Nature* **394**, 840 (1998); L.Vaidman and N.Yoran, *Phys. Rev. A*, to be published, quant-ph/9809063.
- [91] For a comprehensive review see, for example, A.Steane, *Rep. Prog. Phys.* **61**, 117 (1998).
- [92] D.Home and R.Chattopadhyaya, *Phys. Rev. Lett.* **76**, 2836 (1996); see also quant-ph/9903036.

- [93] S.M.Roy and V.Singh, *Mod. Phys. Lett. A* **10**, 709 (1995); see also quant-ph/9811041.
- [94] J.S.Bell, in *The Ghost in the Atom*, (P.C.W.Davies and J.R.Brown, eds.) (Cambridge University Press, Cambridge, 1986).





Quantum Field Theory (QFT) is perhaps the single most important concept in physics to be discovered in the twentieth century. This volume reflects the multidimensional impact of QFT on the evolution of physics in the last century. Freeman Dyson, one of the architects of modern QFT, in a foreword to this volume has warmly endorsed this theme. This volume projects the central theme through a selection of invited articles in the areas where the impact of QFT has been especially pronounced, from particle physics to string theory (with several interpolating stages of development), and extending to some facets of astrophysics and the physics of condensed matter.

In the area of particle physics the emphasis is mainly on symmetries, topologies, gauge theories and renormalization groups. While electroweak interactions have been treated with standard rigour, the strong interaction sector has needed greater filtration, so as to conform to the QFT-oriented thrust of this volume.

A distinct feature of this edited volume is that its theme has been highlighted through a comprehensive editorial summary of all the articles, preceding their classified presentation in six distinct parts:

i) basic structure of QFT; ii) topological aspects of QFT; iii) miscellaneous formal methods in QFT; iv) extension of QFT frontiers; v) QFT in 2+1 dimensions; and vi) strong interaction methods in QFT/QCD.

The contributors range from veterans like (the late) Vladimir Gribov, Marcos Moshinsky, Kazuhiko Nishijima, John Schwarz, Dmitri Shirkov and Edward Witten, to a string of acknowledged experts in their respective fields of expertise, all the way to a few young and promising workers.

The wide range of topics covered makes the book more than just an introductory text book on QFT. It is recommended as a reference book for a broad spectrum of readership, from fresh postdoctoral level in most key areas of QFT to the specialists in evolving areas. The freelance researcher in QFT should also find enough appetizers to kindle his interest in this field.

#### About The Cover

Shiva's Cosmic Dance is conceived here as the rhythmic vibration of interacting Quantum Fields, which holds the key to an understanding of the dynamics of the Universe, from the tiny quark to the immense intergalactic space. A flavour of this scenario may be found in Rabindranath Tagore's poem, *Nrityer o taale shundar holo bidhrohi paramanu* (the rebellious atom tamely submitted, as if in a reverie, to the tune of Shiva's rhythmic Dance).